QUANTIFYING MARINE VIRUS-HOST ABUNDANCE RELATIONSHIPS

A Dissertation
Presented to
The Academic Faculty

By

Charles H. Wigington

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biological Sciences

Georgia Institute of Technology

December 2017

QUANTIFYING MARINE VIRUS-HOST ABUNDANCE RELATIONSHIPS

Approved by:

Dr. Joshua S. Weitz, Advisor
School of Biological Sciences
School of Physics
Georgia Institute of Technology

Dr. Sam Brown
School of Biological Sciences
Georgia Institute of Technology

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Peng Qiu
School of Biomedical Engineering
Georgia Institute of Technology

Dr. Frank Stewart
School of Biological Sciences
Georgia Institute of Technology

Date Approved: August 15, 2017

To those who undertake to understand marine viruses and their hosts.

May your efforts be fruitful and your successes great.

ACKNOWLEDGEMENTS

I would like to thank everyone who had a hand in my success.

My thesis committee including Professors Sam Brown, Frank Stewart, King Jordan, and Peng Qiu all have my most sincere gratitude.

The Weitz Group and my non-Weitz Group collaborators have been a constant source of inspiration and admiration. Most notably Professor Joshua Weitz provided guidance and mentorship. From post-docs to masters students, everyone in the lab contributed to my development as a scientist including Stephen Beckett, Alex Bucksch, Abhi Das, Adrian Lawsin, Nanditha Rajamani, Joey Leung, Hayriye Gulbudak, Ceyhun Eksin, Gabriel Mitchel, Cesar Flores, Luis Jover, Bradford Taylor, Shengyun Peng, Keith Paarporn, Ashley Coenen, Yui-hui Lin, Devika Singh, Walker Gussler, and Daniel Muratore. A special thanks to Chris Follett and Cael Barry in Mick Follows' lab at M.I.T.

My incomparable wife, Cori, was a constant source of clarity. Our oldest son, Owen, was a constant source of perspective. Our newest addition, Andrew, was a constant source of motivation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Marine microbes are the most abundant cellular life forms on Earth yet we know viruses to be even more numerous than microbes. Marine microbes are estimated to total $\sim 10^{30}$ while marine viruses have long been assumed to be ten-times more abundant, thus roughly 10^{31} viruses are estimated to be present in the global ocean. In this thesis, the numerical relationship between viruses and hosts is examined through the perspective provided by global marine datasets. Fixed models which describe virus densities in terms of microbe densities such as the 10:1 ratio are examined for their predictive capacity, ultimately dispelling their use in favor of a non-linear model referred to here as the power-law model. Not only is the ability to predict virus densities from microbe densities improved by the power-law model but further analyses describe how the virus densities vary across across microbial densities, highlighting that the power-law model is most reliable when predicting virus densities of low-microbe density samples. Potential causes of the variability of virus densities are examined, ultimately determining that the studies which collected the samples are themselves a non-trivial source of VMR variability. The data examined in this thesis comes from 22 different studies, totalling more than 5,500 records, and presents an opportunity to test the current knowledge of virus to microbe ratios with empirical data. Finally, the methods and data used in this analyses are described in detail and provided freely to the public to use to so that the findings presented in this thesis can be easily expanded upon by less quantitative but otherwise dedicated researchers in the marine microbiology community.

CHAPTER 1
INTRODUCTION

1.1 Historical context

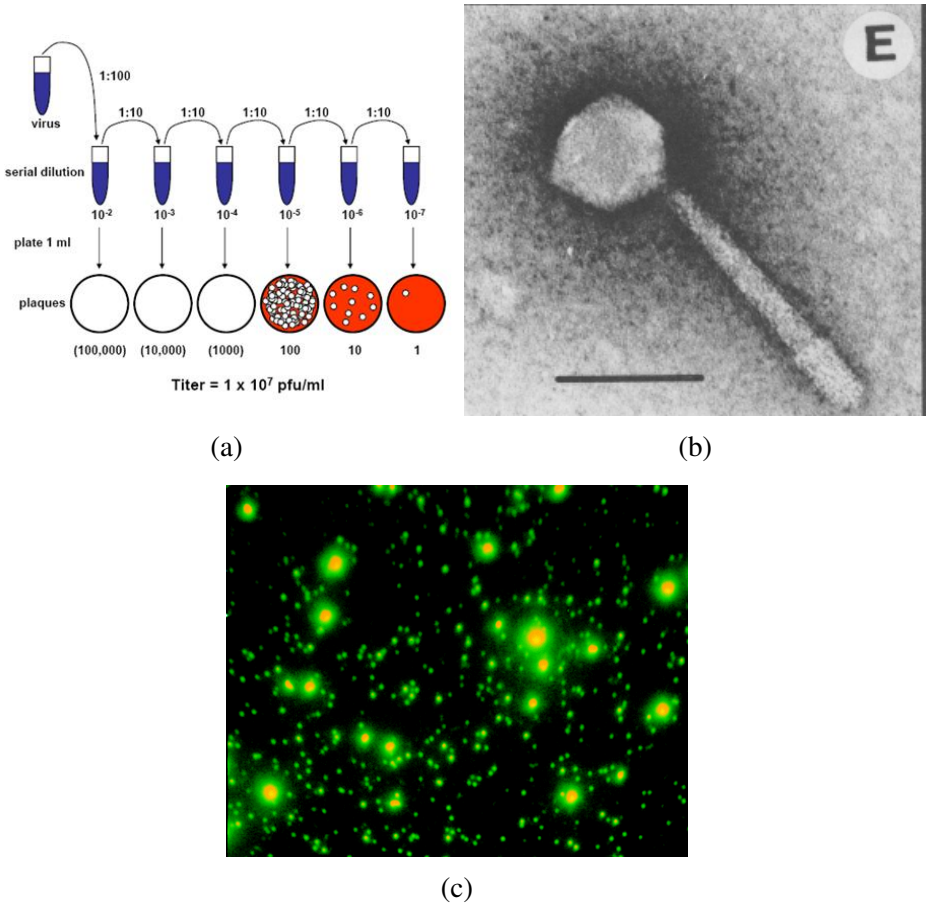
Viruses which require microbes as their hosts were first identified in 1915 by the English microbiologist Frederick Twort and again in 1917 by the French-Canadian microbiologist Felix d'Hérelle. Their discoveries occurred independently of one another and while Twort sought to understand why microbial growth on a surface covered by microbes was cleared of bacteria in spots[1], d'Hérelle sought to understand the mechanism by which successive *Shigella dysenteriae* cultures were cleared of the bacteria when inoculated by material from a previously cleared sample[2]. Shortly thereafter in 1925 Arloing reported that bacteriophage occur naturally in oceanic environments but it wasn't until the work of Zobell in 1946 and Kriss and Rukina in 1947 that marine bacteriophage were described in great detail[3]. The low densities of bacteriophage in marine samples relagated the role of viruses in marine environments to something of a curiosity; although present, they were largely unappreciated in their cumulative effect until the work of Torrella and Morita in 1979 and Bergh ten years later which showed that virus densities in marine environments had been significantly underestimated by multiple orders of magnitude for decades[4, 5]. The discovery that bacteriophage are highly abundant in marine environments spurred research efforts to understand the role of marine viruses in the environment utlimately leading to an understanding that marine viruses are major players in driving marine bacteria evolution, host diversity, and biogeochemical cycling in the ocean.

1.2 Methods for calculating virus to microbe ratios

The methods used to identify and count marine bacteriophage have come along way from the culture-based methods in the early days of bacteriophage research. When bacteriophage were first identifed, viruses were isolated by culture-based techniques meaning that not only did a microbe capable of being infected need to be used to host the bacteriophage but the same microbe type of microbe must be capable of being cultured. To identify bacteriophage via culture-based methods, an agar plate is covered with susceptible microbes and inocculated by a virus-containing solution. Microbes which are infected and lysed by bacteriophage leave clear spots (”watery spots” according to Twort) where the microbes on the surface of the agar have been cleared. The number of bare spots (also called plaques or clearings) is indicative of the density of viruses present in the solution used for inocculation. This process of serial diluation is shown in the left panel of figure 1.1[6]. Culture-based approaches are both time- and labor-intensive and can be biased, i.e., those microbes and viruses that can be cultured may not reflect the most abundant types in the environment.

Using transmission electron microscopy (TEM) in 1979, Torrella and Morita described the density of marine viruses as being several orders of magnitude greater than previously observed [4]. TEM uses a beam of electrons instead of light to create an extremely high resolution image of a sample. One such image taken by Torrella and Morita can be seen in the panel on the right in figure 1.1[6]. However the greater precision with which virus densities in marine environmens could be enumerated comes with a cost: processing samples by TEM is a time- and labor- intensive task. The time-, labor-, and skill- overhead required to make electron micrographs left the door open for the adoption of other methods to count viruses, one of which is currently in use today.

Today virus counting is largely accomplished by epifluorescent microscopy, an example of the result of which can be seen in figure 1.1 taken by Fuhrman in 1999[7]. First stains are bound to genetic material in a sample and are illuminated by a special light. The stain



(a) (b)

(c)

Figure 1.1: **Methods of counting viruses in marine environments have changed over time.**(a) Schematic of serial dilution process taken from [6], (b) electron micrograph of TEM of marine virus from [4], (c) fluorescing viruses and microbes visible as green light as seen in micrograph from [7].

on the sample fluoresces (emits light) allowing for counting of viruses and microbes. The counting process has been automated by the inclusion of image processing software thus the precision and reproducibility of counting viruses has become standardized and is the most reproducible among the methods available today.

1.3 Virus-to-microbe ratios as a target for theory

Bergh reported in 1989 that virus levels in unpolluted seawater are as high as 10^8 per mL, noting that viruses in marine environments have been under-counted and therefore under-appreciated in their role in marine processes[8]. The subsequent outpouring of research

to characterize the levels of marine virus viruses in the ocean has lead to the adoption of ecological models that include virus-host interactions[9, 10, 11]. Evaluation of dynamic ecological models often uses the ratio of viruses to hosts to assess the suitability of underlying mechanisms. Over the past 30 years, describing viruses as ten time more abundant that microbes has become a *de facto* standard in both empirical and theoretical practice.

The ratio of viruses to microbes in the ocean is not always 10:1; virus-to-microbe ratios (VMR) reflect two emormous population sizes each fluctuating over the course of hours thus VMR varies by time even when samples are otherwise identical. An anlysis by Parikka to describe VMR across environments concludes that VMR values observed in marine environments are different from VMR value observed in soil or ocean sediment[12]. Parikka further notes that VMR can be a poor metric to describe an environment beause it is a ratio of two populations. Two environments with equal VMR values should not be considered to be the same simply because of an equal VMR value. In fact, the environments from which two equal VMR values are found can be very different because VMR fails to capture the biological processes in each environment. For example, one explanation for low VMR values comes from Knowles who proposes the "piggyback-the-winner" hypothesis which suggests temperate dynamics become more important in environments densely packed with microbes, i.e. an increase in lysogeny causes declining VMR values.[11] Yet alternative explanations for low VMR values exist. Weitz et al. counter Knowles' argument saying that the assertion that viruses switch strategies in high host density environments, thus explaining the relationship between VMR and microbe densities, is insufficient. "...Many dynamical models of virus-microbe interactions can yield nonlinear relationships between virus and microbial cell density when parameter variation is considered." [13] Thus viruses changing infection strategies and microbes successfully evading infection, both very different biological processes, could generate the same VMR value.

1.4 The contribution of this work to the state of the art

The work presented in this thesis examines three areas: marine virus-to-microbe ratios, models of the relationship between viruses and microbes, and factors which are thought to impact microbial VMR.

First, the often asserted 10:1 ratio of viruses to microbes was examined in the face of empirical data. Simply multiplying microbe-density by ten to yield virus density is challenged by examining the predictive value of this process followed by examining alternative methods to yield estimates of virus densities. Improved estimation of virus densities around the world provides an update to our understanding of viral titres relative to microbe titres in nature. Quantitative analysis of VMR also sets a baseline for characterizing typical variation found in marine environments. Treated as a linear model relating microbe density per mL to virus density per mL, the predictive capacity of the fixed 10:1 ratio was compared to other fixed ratio models. The results of both near-surface ($\leq$ 100m) and sub-surface (100m) analyses indicate that fixed ratios are poor predictors of virus density. In fact, ignoring microbe density entirely and simply asserting there are 17,282,903 viruses in one milliliter of near-surface seawater and 1,804,612 viruses in one milliliter of sub-surface seawater would be more accurate that using the 10:1 ratio.

Allowing the relationship between viruses and microbes to be described as nonlinear resulted in a greater predictive capacity than was seen by any fixed ratio model. In fact, this relationship is not only nonlinear but sub-linear meaning that when looking at seawater sampled in order of increasing microbial density, the ratio of viruses to microbes decreases.

Second, when the virus and microbe densities of seawater samples are visualized in log-log virus-microbe density space, one could argue that variability in virus densities is observed to increase with increasing microbial density, i.e. in high-microbe density samples virus densities vary greater than in low-microbe density samples. Based on visual inspection alone, one could also take the opposite position: virus density variability is ef-

fectively constant across all samples regardless of whether the sample is a high- or low-microbial density sample. Both models are evaluaed using data from around the world, identifying the variable-variance model as the one which fit the data best. The source of the increasing virus density variability is examined and is shown to come from differences in the studies which provided the data, possibly resulting from differences in the way in which data was collected and processed before this analysis. Surprisingly, the variability of virus-to-microbe ratios of one study stood out from the group, thus roughly the same variability in the ratio of viruses to microbes is experienced across studies even though median virus-to-microbe ratios by studies span nearly two orders of magnitude.

Finally, because 22 studies from around the world contributed data to this analysis, each record was given a label which clustered the data according to the geographic location where the sample was taken. Qualitative differences in the virus-host relationship became obvious when examined according to this labeling scheme. Some virus-host relationships were positive, some were negative, and others were neutral. This finding indicates that the labeling of studies by areas expected to be highly biochemically similar was undermined by the cohesiveness of the data within the study which contributed the records to this analysis. Obvious differences in the relationship between viruses and microbes are evident across studies even within the same environment indicating that systematic biases exist in the data, either stemming from low reproducibility across studies or even greater division exists in the labeling of data according to the biochemical similarity of the environment.

CHAPTER 2
RE-EXAMINATION OF THE RELATIONSHIP BETWEEN MARINE VIRUS AND MICROBIAL CELL ABUNDANCES

Adapted from Charles H. Wigington, Derek Sonderegger, Corina P. D. Brussaard, Alison Buchan, Jan F. Finke, Jed A. Fuhrman, Jay T. Lennon, Mathias Middelboe, Curtis A. Suttle, Charles Stock, William H. Wilson, K. Eric Wommack, Steven W. Wilhelm & Joshua S. Weitz. Re-examination of the relationship between marine virus and microbial cell abundances Nature Microbiology 1:15024. doi:10.1038/nmicrobiol.2015.24(2016)
The correction to the published analysis (in review) removes 3 studies (163 records) which were incorrectly assigned marine longitude and latitude coordinates and is reflected in this analysis.

2.1 Abstract

Marine viruses are critical drivers of ocean biogeochemistry and their abundances vary spatiotemporally in the global oceans, with upper estimates exceeding 10^8 per ml. Over many years, a consensus has emerged that virus abundances are typically 10-fold higher than microbial cell abundances. However, the true explanatory power of a linear relationship and its robustness across diverse ocean environments is unclear. Here, we compile 5,508 microbial cell and virus abundance estimates from 22 distinct marine surveys and find substantial variation in the virus-to-microbial cell ratio (VMR), in which a 10:1 model has either limited or no explanatory power. Instead, virus abundances are better described as nonlinear, power-law functions of microbial cell abundances. The fitted, scaling exponents are typically less than 1, implying that VMR decreases with microbial cell density, rather than remaining fixed. Observed scaling also implies that viral effect sizes derived from "representative" abundances require substantial refinement to be extrapolated to regional

or global scales.

2.2 Introduction

Viruses of microbes have been linked to central processes across the global oceans, including biogeochemical cycling [10, 14, 15, 16, 17, 18] and the maintenance and generation of microbial diversity [10, 19, 16, 20, 21]. Virus propagation requires that virus particles both contact and subsequently infect cells. The per cell rate at which microbial cells – including bacteria, archaea, and microeukaryotes – are contacted by viruses is assumed to be proportional to the product of virus and microbial abundances [22]. If virus and microbe abundances were related in a predictable way it would be possible to infer the rate of virus-cell contacts from estimates of microbial abundance alone.

Virus ecology underwent a transformation in the late 1980s with the recognition that virus abundances, as estimated using culture-independent methods, were orders of magnitude higher than estimates via culture-based methods [5]. Soon thereafter, researchers began to report the "virus to bacterium ratio" (VBR) as a statistical proxy for the strength of the relationship between viruses and their potential hosts in both freshwater and marine systems [23]. This ratio is more appropriately termed the "virus-to-microbial cell ratio" (VMR) – a convention which we use here (see Supplementary Text 1).

Observations accumulating over the past 25 years have observed wide variation in VMR, yet there is a concensus that a suitable first-approximation is that VMR is 10 (see Table S1). This ratio also reflects a consensus that typical microbial abundances are approximately 10^6 per ml and typical virus abundances are approximately 10^7 per ml [24, 25]. Yet, the use of a fixed ratio carries with it another assumption: that of linearity, i.e., if microbial abundance were to double, then viruses are expected to double as well. An alternative is that the relationship between virus and microbial abundance is better described in terms of a nonlinear relationship, e.g., a power-law.

In practice, efforts to predict the regional or global-scale effects of viruses on marine

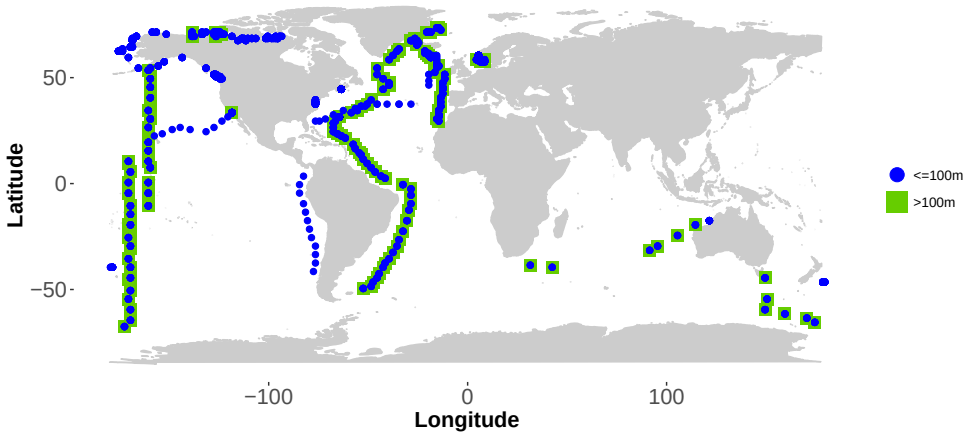


Figure 2.1: **Global distribution and information regarding sample sites.** Each point denotes a location from which one or more samples were taken. Samples range from the surface to up to 5,500 meters below sea level, with 2,758 taken near the surface ($\leq$100 m), noted as squares, and 2,750 taken below the surface ($>$100m), noted as circles. The number of points for each study – the "Frequency" – is found in Table S2.

microbial mortality, turnover and even biogeochemical cycles, depend critically on the predictability of the relative density of viruses and microbial cells. The expected community-scale contact rate, as inferred from the product of virus and microbial abundances, represents a key factor for inferring virus-induced cell lysis rates at a site or sites (e.g., [26]) which also depends on diversity [27], latent infections [28], and virus-microbe infection networks [29]. Here, we directly query the nature of the relationship between viruses and microbial densities via a large-scale compilation and re-analysis of abundance data across marine environments.

2.3 Results

2.3.1 VMR exhibits substantial variation in the global oceans

In the compiled marine survey data (see Figure 2.1, Table 3.1 and the Materials and Methods), 95% of microbial abundances range from 9.0×10^3 to 3.2×10^6 per ml and 95% of

virus abundances range from roughly 3.6×10^5 to 6.5×10^7 per ml (Figure 2.2A). Both microbial and virus concentrations generally decrease with depth as reported previously (e.g., see [30]). We separated the samples according to depth using an operational definition of the near-surface and sub-surface, corresponding to samples taken at depths ≤ 100 and > 100 m, respectively. The cutoff of 100 m was chosen as a typical depth scale for the euphotic zone in systems with low to moderate chlorophyll [31]. The precise depth varies spatiotemporally. Our intent was to distinguish zones strongly shaped by active planktonic foodweb dynamics in well-lit waters, i.e., the "near-surface", from dark mesopelagic waters shaped primarily by decaying particle fluxes with greater depth, i.e., the "sub-surface". The median VMR for the near-surface samples (≤ 100 m) is 11.1 and the median VMR for the sub-surface samples (> 100 m) is 16.0. In that sense, the consensus 10:1 ratio does accurately represent the median VMR for the surface data. We also observe substantial variation in VMR, as has been noted in prior surveys and reviews (see Table S1). Figure 2.2B shows that 95% of the variation in VMR in the near-surface ocean lies between 2.6 and 160 and between 3.9 and 74 in the sub-surface ocean. For the near-surface ocean, 50% of the VMR values are between 5 and 15, 13% are less than 5 and 37.5% exceed 15. This wide distribution, both near- and sub-surface demonstrates potential limitations in utilizing the 10:1 VMR, or any fixed ratio, as the basis for a *predictive* model of virus abundance derived from estimates of microbial abundance.

2.3.2 Virus abundance does not vary linearly with microbial abundance

Figure 2.3 shows two alternative, predictive models of the relationship between logarithmically scaled virus and microbial abundances for water column samples. The models correspond to a fixed-ratio model and a power-law model. To clarify the interpretation of fitting in log-log space consider a fixed-ratio model with a 12:1 ratio between virus and

Table 2.1: **Virus and microbial abundance data from 22 different marine virus abundance studies from 11 different lab groups.** A total of 5,508 data points were aggregated. The data collection dates range primarily from 2000 to 2011. Due to sampling convenience, data primarily comes from coastal waters in the northern hemisphere and were collected predominately during the summer months, with the notable exceptions of long-term coastal monthly monitoring sites (USC MO,BATS,Cheasapeake Bay).

Study Name	Lab	Study Type	Location	Regime	Citation
NORTHSEA2001	Bratback	Spatial	North Sea	Coastal	Bratback (Supp. File 1)
RAUNEFJORD2000	Bratback	Temporal	North Sea	Coastal	Bratback (Supp. File 1)
BATS	Breitbart	Temporal	Sargasso Sea	nonCoastal	Parson et al. 2011 [32]
STRATIPHYT1	Brussaard	Spatial	N-Atlantic Transect	nonCoastal	Mojica et al., 2015 [26]
STRATIPHYT2	Brussaard	Spatial	N-Atlantic Transect	nonCoastal	Brussaard (Supp. File 1)
USC MO	Fuhrman	Temporal	Santa Barbara Channel	nonCoastal	Fuhrman et al. 2006 [33]
GEOTRACES	Herndl	Spatial	Atlantic Transect	nonCoastal	de Corte et al. 2012 [34]
GEOTRACES_LEG3	Herndl	Spatial	Atlantic Transect	nonCoastal	Herndl (Supp. File 1)
BEDFORDBASIN	Li	Temporal	North Atlantic Ocean	Coastal	Li and Dickie 2001 [35]
GREENLAND 2012	Middelboe	Spatial	Greenland Sea	nonCoastal	Middelboe(Supp. File 1)
INDIANOCEAN2006	Middelboe	Spatial	Indian Ocean	nonCoastal	Middelboe (Supp. File 1)
KH04-5	Nagata	Spatial	Southern Pacific Ocean	nonCoastal	Yang et al. 2013 [36]
KH05-2	Nagata	Spatial	Northern Pacific Ocean	nonCoastal	Yang et al. 2013 [36]
CASES03-04	Suttle	Spatial	Arctic Ocean	nonCoastal	Payet and Suttle, 2013 [21]
SOG	Suttle	Temporal	Pacific Ocean - Strait of Georgia	Coastal	Clasen et al., 2008 [37]
ARCTICSBI	Wilhelm	Spatial	Gulf of Alaska	Coastal	Balsom, 2003 [38]
FECYCLE1	Wilhelm	Spatial	South Pacific Ocean	nonCoastal	Strzepek et al. 2005 [39]
FECYCLE2	Wilhelm	Spatial	South Pacific Ocean	nonCoastal	Matteson et al. 2012 [40]
NASB2005	Wilhelm	Spatial	North Atlantic Ocean	nonCoastal	Rowe et al. 2008 [41]
POWOW	Wilhelm	Spatial	Pacific Ocean	nonCoastal	Wilhelm (Supp. File 1)
TABASCO	Wilhelm	Spatial	South Pacific Ocean	nonCoastal	Wilhelm et al. 2003 [42]
MOVE	Wommack	Temporal	Atlantic - Chesapeake	Coastal	Wang et al. 2011 [43]

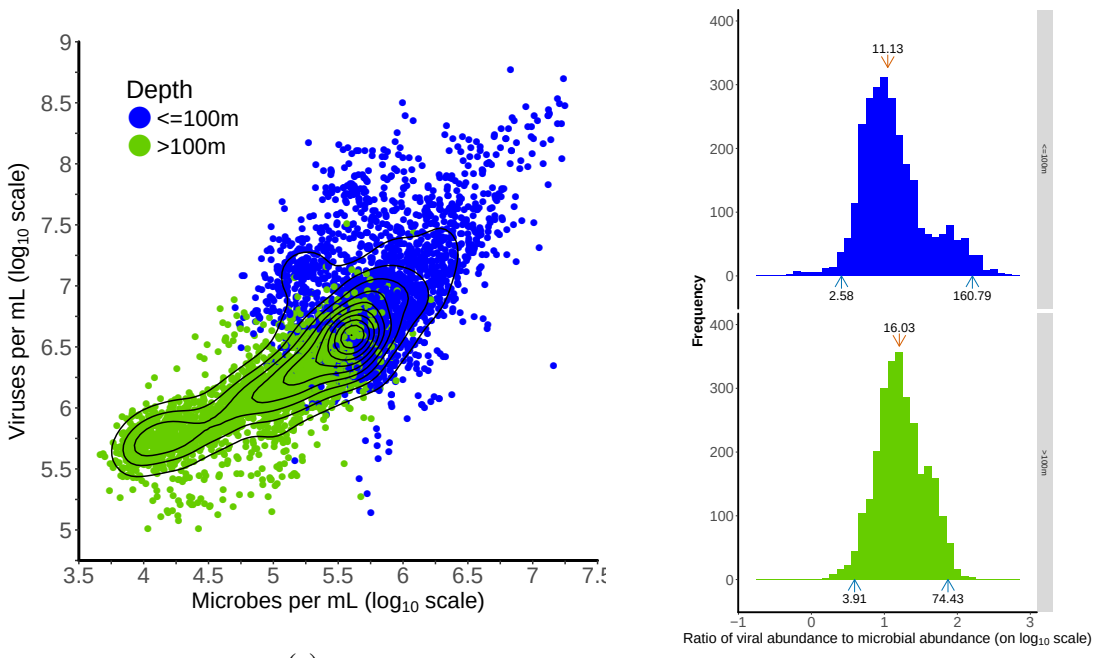
microbial abundance, $V = 12 \times B$. Then, in log-log space the relationship is

$$\log_{10}(V) = \log_{10} 12 + \log_{10} B \tag{2.1}$$

which we interpret as a line with y-intercept of $\log_{10} 12 = 1.08$ and a slope (change in $\log_{10} V$ for a 1-unit change in $\log_{10} B$) of 1. By the same logic, any fixed-ratio model will result in a line with slope 1 in the log-log plot and the y-intercept will vary logarithmically with VMR. The alternative predictive model is that of a power-law: $V = cB^{\alpha_1}$. In log-log space, the relationship is:

$$\log_{10} V = \log_{10} c + \alpha_1 \log_{10} B, \tag{2.2}$$

$$\log_{10} V = \alpha_0 + \alpha_1 \log_{10} B. \tag{2.3}$$



(a) (b)

Figure 2.2: **Variation in virus and microbial abundances and the VMR.** (A) microbial abundance vs. virus abundance colored by depth. Each point represents a biological sample. Contours denote regions of equal probability distribution in the dataset. (B) Histogram of the logarithm of VMR. The top and bottom panels correspond to near- and sub-surface water column samples, respectively. The red arrow denotes the median value and the blue arrows denote the central 95% range of values - where the numbers associated with each arrow denote the non-transformed value of VMR.

The slope, α_1, of a fitted line on log-transformed data denotes the *power-law exponent* that best describes the relationship between the variables. The intercept, α_0, of a fitted line on log-transformed data denotes the logarithmically transformed pre factor.

The 10:1 line has a residual squared error of -6% and -26% in the surface and deep samples, respectively (Table 2). In both cases, this result means that a 10:1 line explains *less* of the variation in virus abundance compared to a model in which virus abundance is predicted by its mean value across the data. In order to evaluate the generality of this result, we considered an ensemble of fixed-ratio models each with a different VMR. In the near-surface samples, we find that only fixed-ratio models between 11.7 and 15.6 have positive R^2 values while outside of this window, all fixed-ratio models explain less of the variation (i.e. have *negative* values of R^2) than does a "model" in which virus abundance is predicted

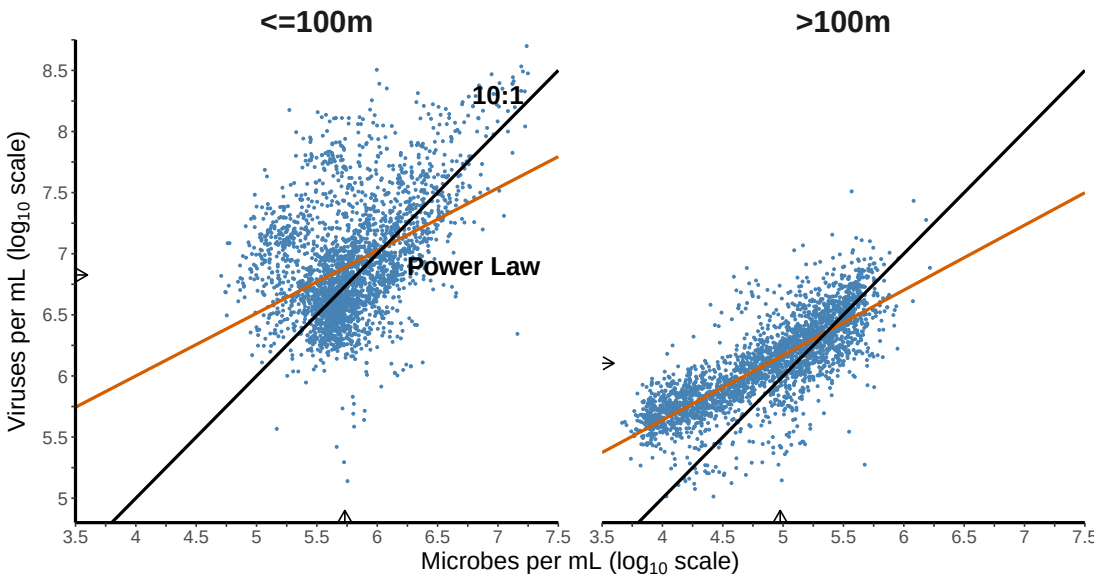


Figure 2.3: **Virus abundance is poorly fit by a model of 10-fold increase relative to microbial abundance.** (Left) Surface ocean – the red line denotes the best fit power-law with an exponent of 0.51 while the black line denotes the 10:1 curve. The best-fit power law explains 19% of the variation and the 10:1 line explains -6% of the variation. See text for interpretation of negative R^2 and the importance of outliers in these fits. (Right) Deeper water column – the red line denotes the best fit power-law with an exponent of 0.53 while the black line denotes the 10:1 curve. The best-fit power law explains 64% of the variation and the 10:1 line explains -26% of the variation. In both cases the arrows on the axes denote the median of the respective abundances.

to be the global mean in the dataset (2.6). This reflects the failure of constant ratio (i.e., linear) models to capture the cluster of high VMRs at low microbial density apparent in the density contours of Figure 2.2A and the shoulder of elevated high VMR frequency in Figure 2.2B. The largest contributor to this cluster of points is the Arctic SBI study (see Table 1). Whereas, in the sub-surface samples, fixed-ratio models in which VMR varies between 12.4 and 23 do have positive explanatory power, but all perform worse than does the power-law model (Figure S1). In contrast, the best fitting power-law model explains 19% and 64% of the variation in the data, for near- and sub-surface samples respectively

(Table 2). The best-fit power-law scaling exponent is 0.51 with 95% confidence intervals (CIs) of $(0.47, 0.55)$ for near-surface samples and 0.53 with 95% CIs of $(0.52, 0.55)$ for sub-surface samples.

The difference between a linear and a power-law model can be understood, in part, by comparing predictions of viral abundances as a function of variation in microbial abundances. For example, doubling microbial abundance along either regression line is not expected to lead to a doubling in virus abundance, but rather a $2^{0.51} = 1.42$ and $2^{0.53} = 1.4$ fold increase, respectively. The difference between models becomes more apparent with scale, e.g., 10- and 100-fold increases in near-surface microbial abundances are predicted to be associated with $10^{0.53} = 3.4$ and $100^{0.53} = 11$ fold increases in viral abundances, respectively, given a power-law model. The power-law model is an improvement over the fixed ratio model in both the near- and sub-surface, even when accounting for the increase in parameters (Table 2). In the near-surface, refitting surface data without outliers improves explanatory power to approximately $R^2 = 0.33$ in contrast to an $R^2 = 0.67$ for the sub-surface (see Methods and Figure S2). Power law exponents in the near- and sub-surface are qualitatively robust to variation in the choice of depth threshold, e.g., as explored over the range between 50 m and 150 m (see Figure S3). In summary, the predictive value of a power-law model is much stronger in the sub-surface than in the near-surface, where confidence in the interpretation of power-law exponents is limited.

2.3.3 Study-to-study measurement variation is unlikely to explain the intrinsic variability of virus abundances in the surface ocean

Next, we explored the possibility that variation in methodologies affected the baseline offset of virus abundance measurements and thereby decreased the explanatory power of predicting virus abundances based on microbial abundances. That is, if V^* is the true and unknown abundance of viruses, then it is possible that two studies would estimate $\hat{V}_1 = V^*(1 + \epsilon_1)$ and $\hat{V}_2 = V^*(1 + \epsilon_2)$ where $|\epsilon_1|$ and $|\epsilon_2|$ denote the relative magnitude

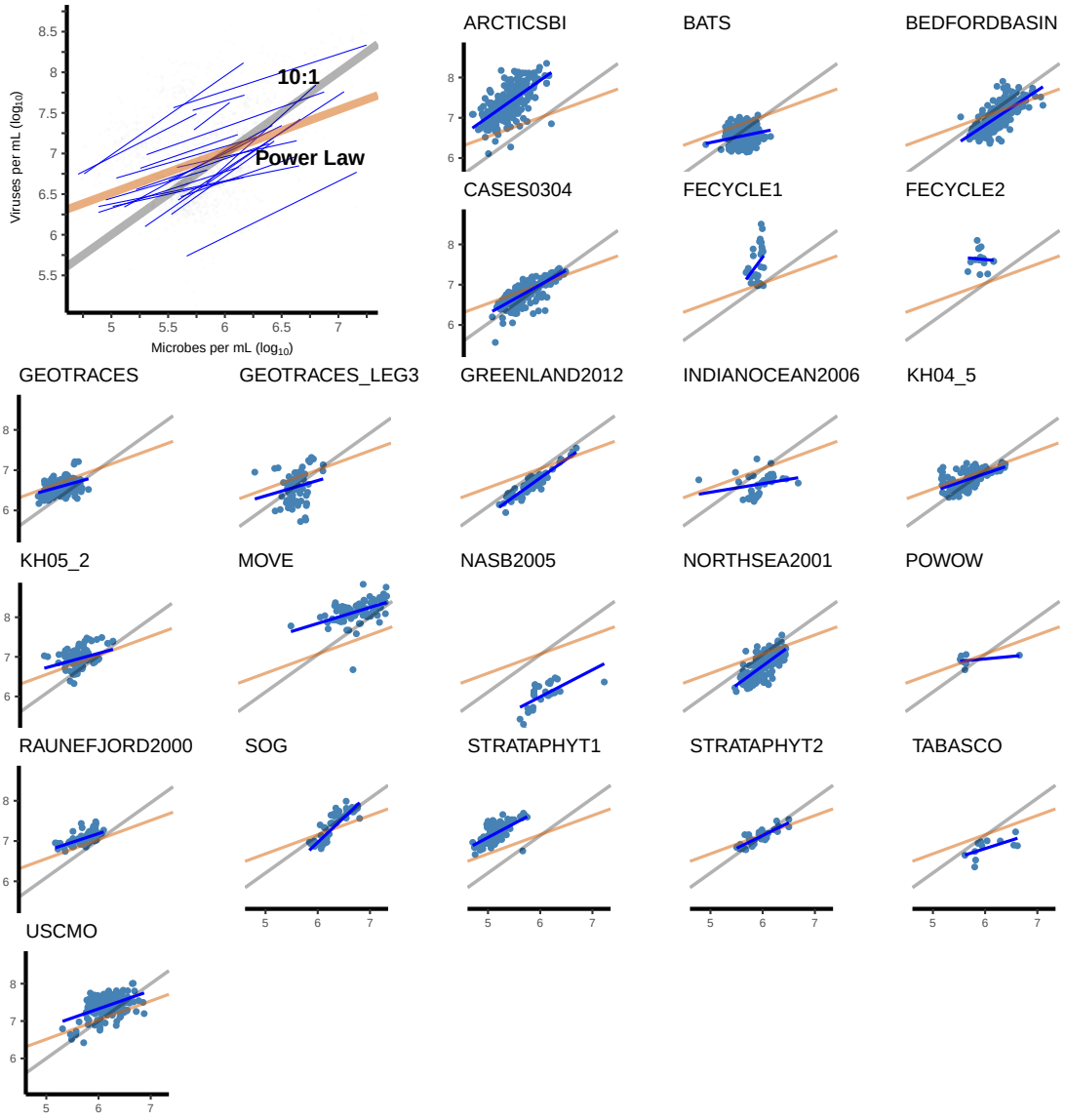


Figure 2.4: **Virus-microbial relationships given the variable slope and intercept mixed-effects model.** (Upper-left) Best-fit power-law for each study (blue lines) plotted along with the best-fit power-law of the entire dataset (red line) and the 10:1 line (grey line). (Individual panels) Best-fit power-law model (blue line) on log-transformed data (blue points) for each study, with the power-law model regression (red) and 10:1 line (black) as reference. The power-law exponents and associated confidence intervals are shown in Figure 2.5.

of study-specific shifts. We constrain the relative variation in measurement, such that the measurement uncertainty is 50% or less (see Materials and Methods). The constrained regression model improves the explanatory power of the model (see Table 2), but in doing so, the model forces 18 of the 22 studies to the maximum level of measurement variation per-

mitted (Figure S4). We do not expect that differences in measurement protocols to explain nearly 2 orders of magnitude variation in estimating virus abundance given the same true virus abundance in a sample. Note that when sub-surface samples were analyzed through the constrained power-law model, there was only a marginal increase of 2% in R^2 and, moreover, 9 of the 12 studies were fit given the maximum level of measurement variation permitted (Figure S4). The constrained intercept model results suggest that the observed variation in virus abundance in the surface oceans is not well explained strictly by variation in measurement protocol between studies.

2.3.4 VMR decreases with increasing microbial abundance – a hallmark of power-law relationships

We next evaluate an ensemble of power-law models: $V_i = c_i N^{\alpha_i}$ where the index i denotes the use of distinct intercepts and power-law exponents for each survey. The interpretation of this model is that the nonlinear nature of the virus to microbial relationship may differ in distinct oceanic realms or due to underlying differences in sites or systems, rather than due to measurement differences. Figure 2.4 shows the results of fitting using the study-specific power-law model in the surface ocean samples. Study-specific power-law fits are significant in 16 of 22 cases in the surface ocean. The median power-law exponent for studies in the surface ocean is 0.48. Furthermore, of those significant power-law fits, the 95% distribution of the power-law exponent excludes a slope of one and is entirely less than one in 9 of 16 cases (see Figure 2.5). This model in which the power-law exponent varies with study is a significant improvement in terms of R^2 (Table 2). For sub-surface samples, study-specific power-law fits are significant in 10 of 12 cases in the sub-surface (Figure S5). The median power-law exponent for studies in the sub-surface is 0.67. Of those significant power-law fits, the central 95% distribution of the power-law exponent is less than one in 6 of 10 cases (see Figure S6). A power-law exponent of less than one means that virus abundance increases less than proportionately given increases in microbial

abundance. This study-specific analysis extends the findings that nonlinear, rather than linear, models are more suitable to describe the relationship between virus and microbial abundances. We find that the dominant trend in both near-surface and sub-surface samples is that VMR decreases as microbial abundance increases. The increased explanatory power by study is stronger for near-surface than for sub-surface samples. This increase in R^2 comes with a caveat: study-specific models do not enable *a priori* predictions of virus abundance given a new environment or sample site, without further effort to disentangle biotic and abiotic factors underlying the different scaling relationships.

2.4 Discussion

Viruses are increasingly considered within efforts to describe the factors controlling marine microbial mortality, productivity, and biogeochemical cycles [44, 15, 45, 16, 46, 47]. Quantitative estimates of virus-induced effects can be measured directly, but are often inferred indirectly, using the relative abundance of viruses to microbials. To do so, there is a consensus that assuming the virus-to-microbial ratio is 10 in the global oceans - despite observed variation - is a reasonable starting point. Here, we have re-analyzed the relationship of virus to microbial abundance in 22 marine survey datasets. We find that 95% of the variation in VMR ranges from 2.6 to 160 in the near-surface ocean and from 3.9 to 74 in the sub-surface. Although the 10:1 ratio accurately describes the median of the VMR in the surface ocean, the broad distribution of VMR implies that microbial abundance is a poor quantitative predictor of virus abundance. Moreover, increases in microbial abundance do not lead to proportionate increases in virus abundance. Instead, we propose that the virus to microbial abundance relationship is nonlinear, and that the degree of nonlinearity – as quantified via a power-law exponent – is typically less than 1. This sublinear relationship can be interpreted to mean that VMR decreases as an increasing function of microbial abundance, and generalizes earlier observations [24].

Power-law relationships between virus and microbial abundance emerge from complex

feedbacks involving both exogeneous and endogenous factors. The question of exogenous factors could be addressed, in part, by examining environmental covariates at survey sites. For example, if microbial and virus abundances varied systematically with another environmental co-factor during a transect, then this would potentially influence the inferred relationship between virus and microbial abundances. In that same way, variation in environmental correlates, including temperature and incident radiation, may directly modify virus life history traits [48, 49]. Likewise, some of the marine survey datasets examined here constitute repeated measurements at the same location (e.g., at the Bermuda Atlantic Time-series Study (BATS)). Time-varying environmental factors could influence the relative abundance of microbes and viruses. It is also interesting to note that viruses-induced mortality is considered to be more important at eutrophic sites [24], where microbial abundance is higher - yet the observed decline in VMR with microbial abundance would suggest the opposite.

It could also be the case that variation in endogenous factors determines total abundances. Endogenous factors can include life history traits of viruses and microbes that determine which hosts are infected by which viruses [29] as well as the quantitative rates of growth, defense and infection. For example, relative strain abundances are predicted to depend on niche differences according to the "Kill-the-Winner" theory which presupposes trade-offs between growth and defense [50, 10]. Similarly, the recent hypothesis of a complementary "King-of-the-Mountain" mechanism suggests that relative abundance relationships may depend on life-history trait differences, even when tradeoffs are not strict [51]. In both examples, total abundances may nonetheless depend on other factors, including the strength of grazing.

The analysis of abundance relationships also requires consideration of variation in time. As is well known, virus-microbe interactions can lead to intrinsic oscillatory dynamics. Indeed, previous observations of a declining relationship between VMR and microbial abundance have been attributed to changing ratios across phytoplankton bloom events, including

possible virus-induced termination of blooms [24]. Similar arguments were proposed in the analysis of tidal sediments [52]. Alternatively, observations of declining VMR with microbial density have been attributed to variation in underlying diversity [53]. Another factor potentially complicating abundance predictions is that episodic events, including induction of lysogenic populations, influence total microbial and viral counts. Varying degrees of lysogenic and co-infection relationships have been measured in marine virus-host systems [54, 25, 28], the consequences of which may differ from that given interactions with lytic viruses, as is commonly the focus of model- and empirical-based studies. Whatsoever the mechanism(s), it is striking that virus abundances in some surveys can be strongly predicted via alternative power-law functions of microbial abundances. Mechanistic models are needed to further elucidate these emergent macroecological patterns and relationships, akin to recent efforts to explain emergent power laws between terrestrial predator and prey [55].

The present analysis separated the abundance data first according to depth and then according to survey as a means to identify different relationships between virus and microbial abundances in the global oceans. The predictive value of total microbial abundance is strong when considering sub-surface samples. In contrast, microbial abundance is not a strong predictor of virus abundance in the near-surface samples, when utilizing linear or nonlinear models. The predictive power of nonlinear models improved substantially in the near-surface when evaluating each marine survey separately. The minimal predictive value of microbial cell abundances for inferring viral abundances in the near-surface when aggregating across all surveys is problematic given that virus-microbe interactions have significant roles in driving microbial mortality and ecosystem functioning [44, 15, 17]. Indeed the aggregation of abundance measurements in terms of total microbial abundances may represent part of the problem.

At a given site and time of sampling, each microbial cell in the community is potentially targeted by a subset of the total viral pool. In moving forward, understanding variation in

virus abundance and its relationship to microbial abundance requires a critical examination of correlations at functionally relevant temporal and spatial scales, i.e., at the scale of interacting pairs of viruses and microbes. These scales will help inform comparisons of virus-microbe contact rates with viral-induced lysis rates, thereby linking abundance and process measurements. We encourage the research community to prioritize examination of these scales of interaction as part of efforts to understand mechanisms underlying nonlinear virus-microbe abundance relationships in the global oceans.

2.5 Materials and Methods

2.5.1 Data source

Marine virus abundance data was aggregated from 22 studies (Table 3.1). A total of 5,508 data points were aggregated. The data collection dates range from 1996 to 2012. Data primarily comes from coastal waters in the northern hemisphere and were collected predominately during the summer months, with the notable exceptions of long-term coastal monthly monitoring sites, i.e., the studies USC MO, BATS, and MOVE.

2.5.2 Data processing

Analyses of the data were performed using R version 3.1.1. Scripts and original data are provided at

https://github.com/cwigington3/VIRBAC_analysis.

2.5.3 Power-law model

A power-law regression model used the $\log_{10}$ of the predictor variable, microbial abundance per mL N, and the $\log_{10}$ of the outcome variable, virus abundance per mL V. The power-law regression was calculated using the equation $\log_{10} V = \alpha_0 + \alpha_1 \log_{10} N$. The α_0 and α_1 parameters were fit via OLS regression to minimize the sum of square error.

2.5.4 Constrained variable-intercept model

The constrained model is a "mixed-effects" regression model using the same predictor and outcome variables, $\log_{10}$ of microbial abundance per mL and the $\log_{10}$ virus abundance per mL, respectively. This model includes study-specific intercepts which were constrained such that the value for any of the intercepts were restricted to one standard error above or below the intercept value taken from the power-law model. The standard error value for this model came from the power-law model. The equation for this model is $V = \alpha_0^{(i)} + \alpha_1 N$, where $\alpha_0^{(i)}$ is the study-specific intercept and and α_1 is the slope common to all studies, N is the predictor variable, and V is the outcome variable.

2.5.5 Variable slope and variable intercept model

A power-law model where the exponent and intercept varied with each study was evaluated using the same predictor variable, $\log_{10}$ microbial abundance per mL, and the same outcome variable, $\log_{10}$ virus abundance per mL. In this model, there was a study-specific α_0 and α_1 and a OLS regression calculated using the equation V = $\alpha_0^{(i)} + \alpha_1^{(i)}$N.

2.5.6 Bootstrapping model confidence intervals

Bootstrap analyses of the power-law model and mixed effects models were conducted to derive 95% confidence intervals surrounding the parameters estimated by the models. For all models the original dataset was sampled with replacement, by study, to arrive at a bootstramp sample dataset, this process was repeated 10,000 times. Distributions for all parameters were generated and the 2.5%, 50%, and 97.5% points were identified from among the 10,000 parameter estimates.

2.5.7 Outlier identification

Outliers in the data were identified by calculating the top and bottom 2% of estimated VMR amongst the entire 5,508 samples. The outliers corresponded to ratios below 1.81

and above 128. Those samples with virus to microbial ratios which fell outside of these bounds were considered outliers. There were 218 outlier samples taken at depths $\leq$100 m and 10 outlier samples taken at depths $>$100 m.

2.5.8 Depth Cutoff Robustness

The cutoff point for which data was partitioned into either the near-surface or the sub-surface was varied from 50 meters to 150 meters in 1 meter increments. For each step, a power law model was evaluated for both the near-surface and the sub-surface.

2.6 Supplemental Figures and Text

2.6.1 Operational definitions of viral and microbial abundances

The operational definitions "near-surface" and "sub-surface" are used to indicate predominantly euphotic and aphotic ocean depths [31]. We use the term virus abundance throughout this manuscript to denote estimates derived from culture-independent methods, including epiflourescence microscopy [56] or flow cytometry [57]. Viruses measured in these methods are generally thought to represent bacteriophage, consistent with the numerical dominance of bacteria in seawater [25]. Yet, currently available methods have potential limitations. For example, ssDNA viruses [58, 59], RNA viruses [60, 61], and giant viruses [62] are under-counted when estimates are made via epiflourescence microscopy with standard DNA based stains.

Table 2.2: **Origins and emerging consensus of the 10:1 ratio of virus abundance to microbial cell abundance in aquatic systems - from freshwater lakes to the global oceans.**

Year	Observation	Reference
1894	Marine bacteria are first discussed by Certes, Fischer and Russell	[63, 64, 65, 66]
1915,1917	Bacteriophage are discovered	[**dHerelle˙1917**, 1]
1925	The presence of bacteriophage in seawater is noted	[67]
1946	ZoBell reports that bacteriophage occur only sporadically and in the littoral zone and concludes there is insufficient evidence for viruses to be considered as key to limiting open ocean bacteria	[68, 69]
1947	The presence of bacteriophage described in the oceans	[3]
1979	Using transmission electronic microscopy, up to 10^4 ml^{-1} bacteriophage particles are observed in coastal water, an observations that sparked the rebirth of virus ecology a decade later.	[4]
1989	"Rebirth" of virus ecology across a series off papers begins with a report of virus and bacteria abundances for which VMRs range from 0.2 (Raunefjorden) to 50 (North Atlantic)	[5]
1990	Report of virus particles ranging from 10^6 - 10^{11} per liter, infecting up to 7% of heterotrophic bacteria and each infected cell containing 10-100 mature virions	[70]
1991-1993	Estimates of virus abundance exceeding bacteria abundance by 5-10 fold from a series of papers (this observation noted in [71])	[72, 73, 74, 75, 76]
1995	Maranger and Bird [23] survey 22 Quebec lakes and collect literature from 14 studies [5, 77, 72, 78, 79, 74, 80, 75, 76, 81, 82] and report VMR higher in freshwater (20-25) than marine systems (1-5).	[23]
2000	Wommack and Colwell suggest that VMR typically ranges between 3 and 10, and note that VMR decreases as microbial abundance increases.	[24]
2000	A VMR "roughly equal to 10" (attributed to [23] is designated as a target for parameterizing the Kill-the-Winner theory of virus-microbe interactions.	[10]
2004	Consistency in VMR is attributed to the idea that most viruses are phage that infect bacteria. Notes a VMR of 10 in marine systems and attributes to [23].	[25]
2004	Chibani-Chennoufi and colleagues advance the notion that VMR is 10:1 in the ocean and that this is justified by the claim that each bacterial species can be infected by 10 different phage.	[83]
2008	VMR ratios reviewed in several publications that collate information from multiple studies, with a 10:1 consensus despite noted variation.	[37, 84]
2011	VMR reviewed across several regimes, with evidence for a linear relationship between viruses and microbes in the water column and a nonlinear relationship in sediment.	[30]
2014	The BioNumbers database, intended to facilitate quantitative analysis in the biosciences, lists VMR as 10.	[85]

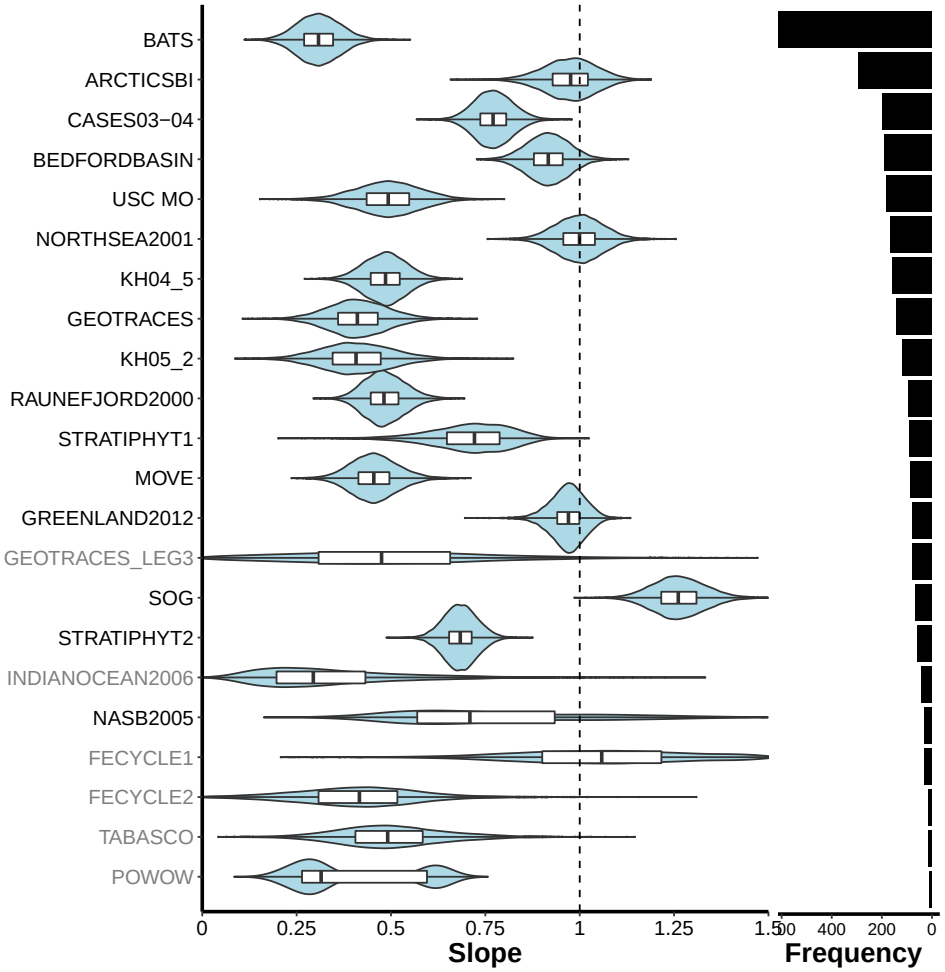


Figure 2.5: **Study-specific 95% confidence intervals of power-law exponents for relationships between virus and microbial abundance in the surface.** The confidence intervals are plotting using "violin" plots including the median (center black line), 75% distribution (white bars) and 95% distribution (black line), with the distribution overlaid (blue shaded area). The number of points included as part of each study is displayed on the right-most bar plots. Study labels in black indicate those studies for which the regression fit had a p-value less than 0.002=0.05/22 (accounting for a multiple comparison correction given the analysis of 22 studies). Study labels in gray indicate a p-value above this threshold.

Table 2.3: **Number of data points per study.**

Study	$\leq$ 100m	$>$ 100m	Total
ARCTICSBI	292	0	292
BATS	626	756	1382
BEDFORDBASIN	188	0	188
CASES03-04	199	46	245
FECYCLE1	31	0	31
FECYCLE2	15	0	15
GEOTRACES	141	631	772
GEOTRACES_LEG3	78	351	429
GREENLAND2012	78	46	124
INDIANOCEAN2006	42	10	52
KH04_5	159	383	542
KH05_2	117	238	355
MOVE	84	0	84
NASB2005	31	0	31
NORTHSEA2001	164	27	191
POWOW	9	0	9
RAUNEFJORD2000	95	0	95
SOG	67	0	67
STRATIPHYT1	89	24	113
STRATIPHYT2	59	34	93
TABASCO	12	0	12
USC MO	182	204	386
Total	2,758	2,750	5,508

Table 2.4: **Information theoretic comparison of alternative models of the relationship between virus and microbial cell abundances.** The values of the Aikake Information Criteria (AIC) are defined in the Materials and Materials and Methods. The value of R^2 for each model denotes the relative amount of variance explained. Negative values of R^2 mean that a model explains less variance than does the overall mean.

	$\leq$ 100m		$>$100m	
Model	R^2	AIC	R^2	AIC
10:1	-0.06	-14637.70	-0.26	-14492.09
Power Law	0.19	-15462.62	0.64	-18313.82
Constrained Power Law	0.44	-16511.94	0.66	-18513.48
Power Law by Study	0.79	-19210.57	0.72	-18972.81

Table 2.5: **Variation in the estimate of the intercept, $\alpha_0^{(i)}$, for each study and associated standard error for the constrained power-law model as applied to surface ocean data.** The common intercept in this model is $\alpha_0 = 3.95$ and the common slope is 0.51. The group column denotes whether the study-specific intercept exceeds that of the common intercept (denoted as group A) or is below that of the common intercept (denoted as group B). The table is sorted according to the lab-specific intercept estimates.

Study	Intercept	Std. Error	Group
ARCTICSBI	4.07	0.16	A
FECYCLE1	4.07	0.19	A
FECYCLE2	4.07	0.20	A
MOVE	4.07	0.20	A
RAUNEFJORD	4.07	0.17	A
STRATAPHYT1	4.07	0.16	A
USC MO	4.07	0.18	A
KH05_2	4.00	0.17	A
SOG	3.92	0.19	B
POWOW	3.90	0.21	B
STRATAPHYT2	3.86	0.18	B
BATS	3.83	0.17	B
BEDFORDBASIN	3.83	0.19	B
CASES0304	3.83	0.17	B
GEOTRACES	3.83	0.16	B
GEOTRACES_LEG3	3.83	0.17	B
GREENLAND2012	3.83	0.18	B
INDIANOCEAN2006	3.83	0.18	B
KHO4_5	3.83	0.17	B
NASB2005	3.83	0.19	B
NORTHSEA2001	3.83	0.18	B
TABASCO	3.83	0.21	B

Table 2.6: **Explanatory power and significance of power-law fits for the model in which the power-law exponent is allowed to vary between studies.** Empty cells in a row denote the absence of samples collected at depths > 100 meters for the study denoted in the left-most column.

	≤ 100 m		> 100 m	
Study	R^2	p-value	R^2	p-value
ARCTICSBI	0.441	<1e-05		
BATS	0.045	<1e-05	0.504	<1e-05
BEDFORDBASIN	0.537	<1e-05		
CASES03-04	0.541	<1e-05	0.072	0.0718
FECYCLE	0.146	0.0341		
FECYCLE2	0.004	0.813		
GEOTRACES	0.163	<1e-05	0.706	<1e-05
GEOTRACES_LEG3	0.043	0.0695	0.396	<1e-05
GREENLAND2012	0.868	<1e-05	0.333	2.7e-05
INDIANOCEAN2006	0.068	0.0955	0.288	0.11
KH04_5	0.325	<1e-05	0.703	<1e-05
KH05_2	0.122	0.000112	0.836	<1e-05
MOVE	0.24	<1e-05		
NASB2005	0.382	0.00021		
NORTHSEA2001	0.542	<1e-05	0.51	2.85e-05
POWOW	0.136	0.329		
RAUNEFJORD2000	0.349	<1e-05		
SOG	0.788	<1e-05		
STRATIPHYT1	0.448	<1e-05	0.471	0.000214
STRATIPHYT2	0.768	<1e-05	0.731	<1e-05
TABASCO	0.371	0.0354		
USC MO	0.229	<1e-05	0.462	<1e-05

Table 2.7: **Power-law exponents, α_1, and intercepts, α_0, for each study from the mixed model allowing study-specific slopes and intercepts.** Empty cells in a row denote the absence of samples collected at depths $>$ 100 m for the study denoted in the left-most column.

	$\leq$ 100 m		$>$ 100 m	
Study	α_0	α_1	α_0	α_1
ARCTICSBI	2.12	0.97		
BATS	4.80	0.31	2.49	0.72
BEDFORDBASIN	1.32	0.91		
CASES03-04	2.40	0.77	2.80	0.65
FECYCLE	1.16	1.07		
FECYCLE2	5.18	0.41		
GEOTRACES	4.39	0.41	3.63	0.52
GEOTRACES_LEG3	4.08	0.45	3.43	0.53
GREENLAND2012	0.97	0.97	2.05	0.76
INDIANOCEAN2006	4.96	0.28	2.75	0.66
KH04_5	4.04	0.48	3.00	0.64
KH05_2	4.65	0.41	2.48	0.76
MOVE	5.05	0.45		
NASB2005	1.83	0.69		
NORTHSEA2001	0.74	1.00	1.59	0.84
POWOW	5.09	0.31		
RAUNEFJORD2000	4.28	0.48		
SOG	-0.67	1.25		
STRATIPHYT1	3.39	0.71	4.53	0.45
STRATIPHYT2	2.92	0.68	2.96	0.67
TABASCO	3.73	0.49		
USC MO	4.37	0.49	2.18	0.79

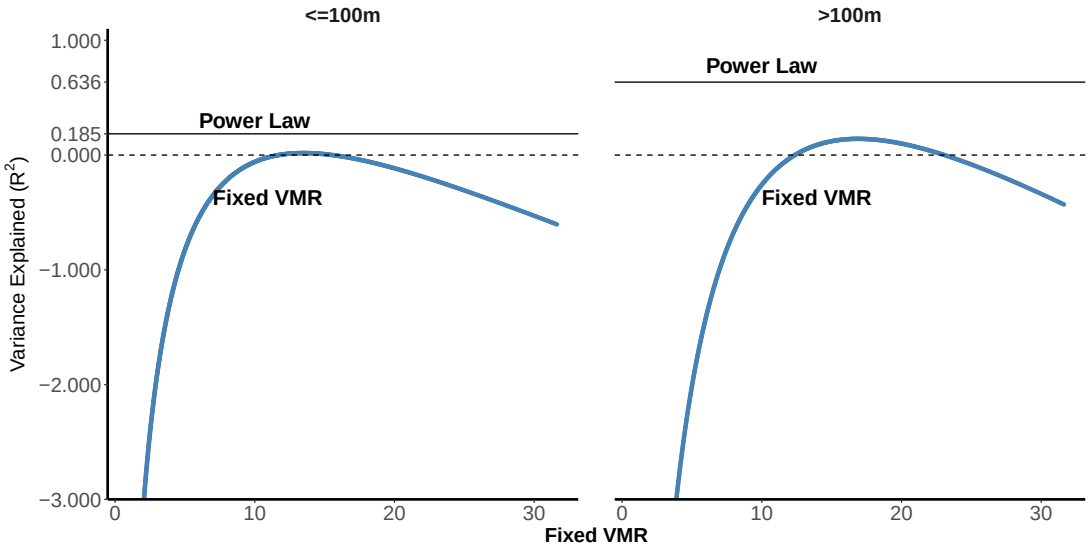


Figure 2.6: **Explanatory power of fixed VMR models in the surface ocean (left) and deeper water column (right).** The x-axis denotes the value r in the model $V = rM$ where V denotes virus abundance and M denotes microbial abundance. The y-axis denotes the fraction of variance explained, R^2. Here, $R^2 = 1 - \mathrm{SSE}_{\mathrm{model}}/\mathrm{SSE}_{\mathrm{total}}$ where $\mathrm{SSE}_{\mathrm{model}}$ is the sum of squared errors for the model and $\mathrm{SSE}_{\mathrm{total}}$ is the sum of total squared errors.

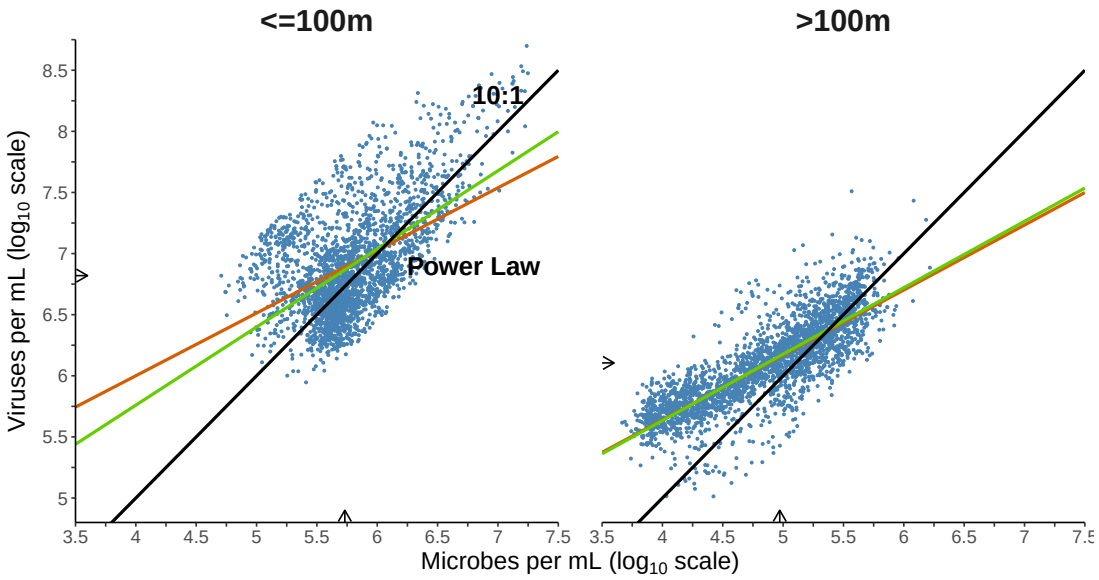


Figure 2.7: **Explanatory power of fixed VMR models in the near-surface and sub-surface with and without outliers.** The three lines in each panel denote the 10:1 line (black), power-law fit (red) and power-law fit when removing outliers (green). The R^2 value for the power law fit for surface data excluding outliers is 0.30, has a slope of 0.58 and an intercept of 3.50. The R^2 value for the power law fit for sub-surface data excluding outliers is 0.65, has a slope of 0.54 and an intercept of 3.49.

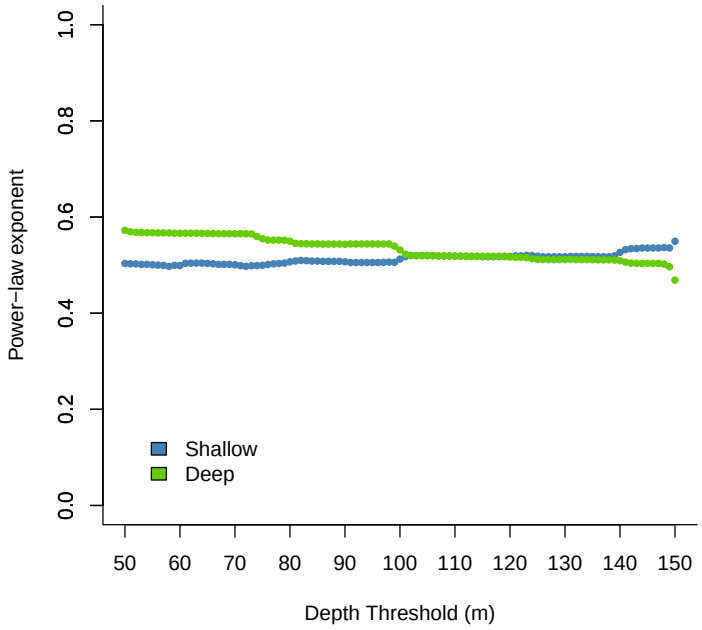


Figure 2.8: **Variation in estimated power-law exponent as a function of sampling depth cutoff, over the range 50m to 150m.** In all cases power-law exponents were measured on log transformed data (see Materials and Methods). The slope varies from $0.40 - 0.47$ for near-surface samples, as compared to the CI of $0.39 - 0.46$ when using 100m cutoffs, i.e., nearly coinciding with the original uncertainty in the estimated slope. The slope varies from $0.47 - 0.57$ for sub-surface samples, as compared to the CI of $0.52 - 0.55$ when using 100m cutoffs. This represents an approximately 10% change in slope estimate. The trend in slope with changes in cutoff depth reflects the difference between near- and sub-surface scaling relationships which are shallower and steeper, respectively. Irrespective of cutoff, we conclude that power-law exponents are sublinear, close to that when estimated using a 100m cutoff.

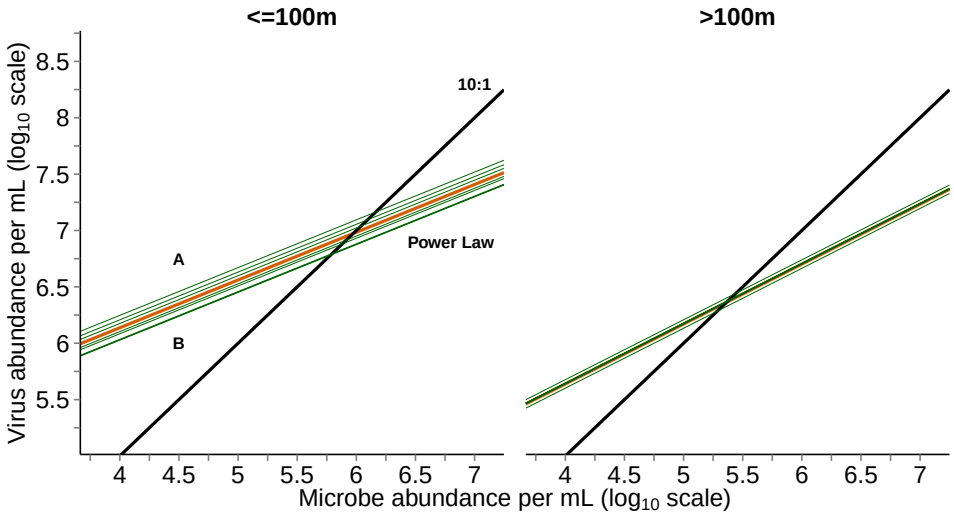


Figure 2.9: **Constrained regression model for samples taken at depths $\leq$ 100m (left) and $>$ 100m (right) where the intercept for each study was permitted to vary (see Materials and Methods)**. Blue line denotes the 10:1 relationships, the red line denotes the best-fitting power-law model, and the remainder of lines denote the variable intercept model with intercept values reported in Table 2.5.

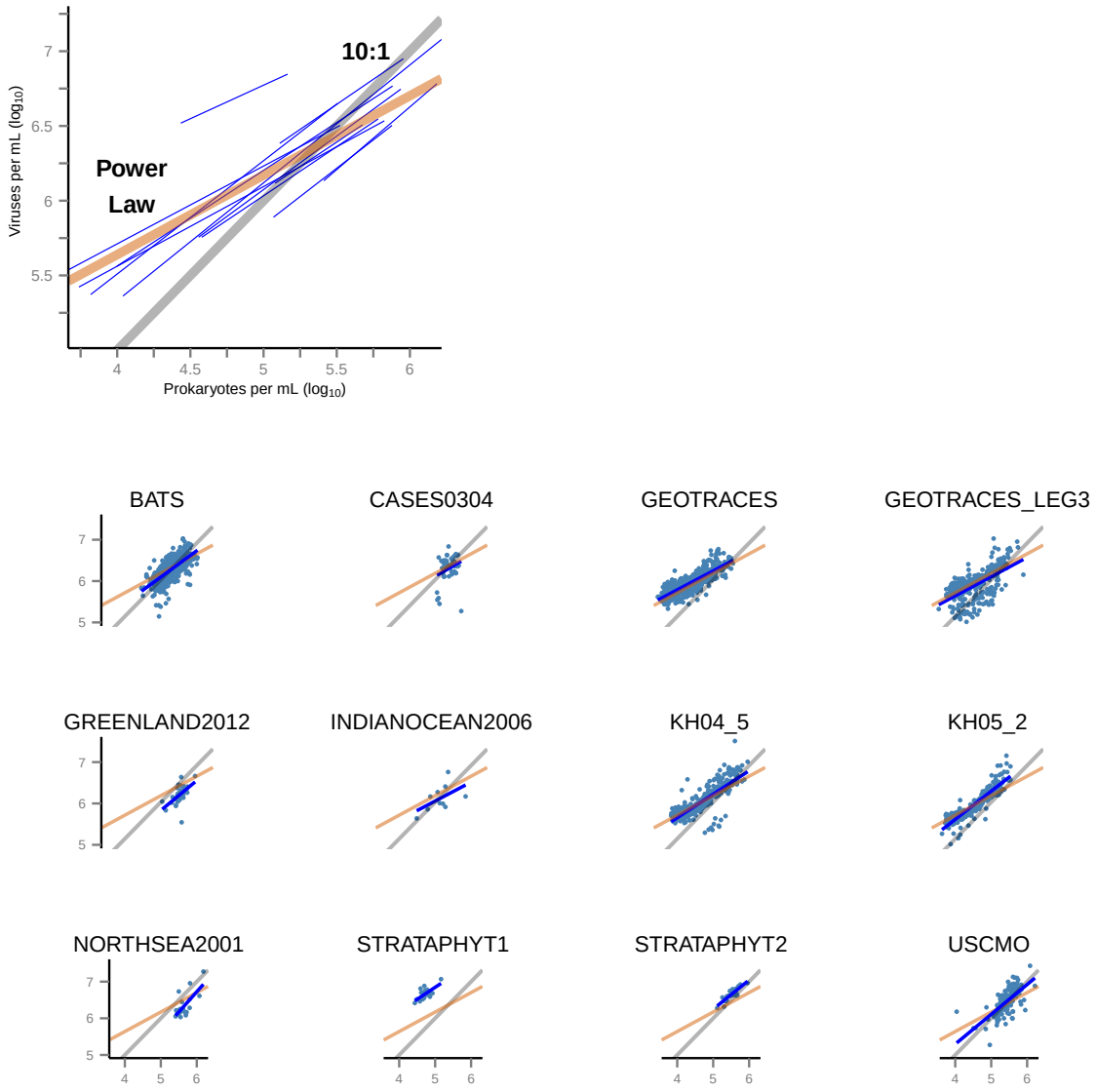


Figure 2.10: **Virus-microbe relationships given the variable slope and intercept mixed-effects model for samples taken at depths greater than 100m.** (Upper-left) Best-fit power-law for each study (blue lines) plotted along with the best-fit power-law of the entire dataset (red line) and the 10:1 line (grey line). (Individual panels) Best-fit power-law model (blue line) on log-transformed data (blue points) for each study, with the power-law model regression (red) and 10:1 line (black) as reference. The power-law exponents and associated confidence intervals are shown in Figure 2.11,

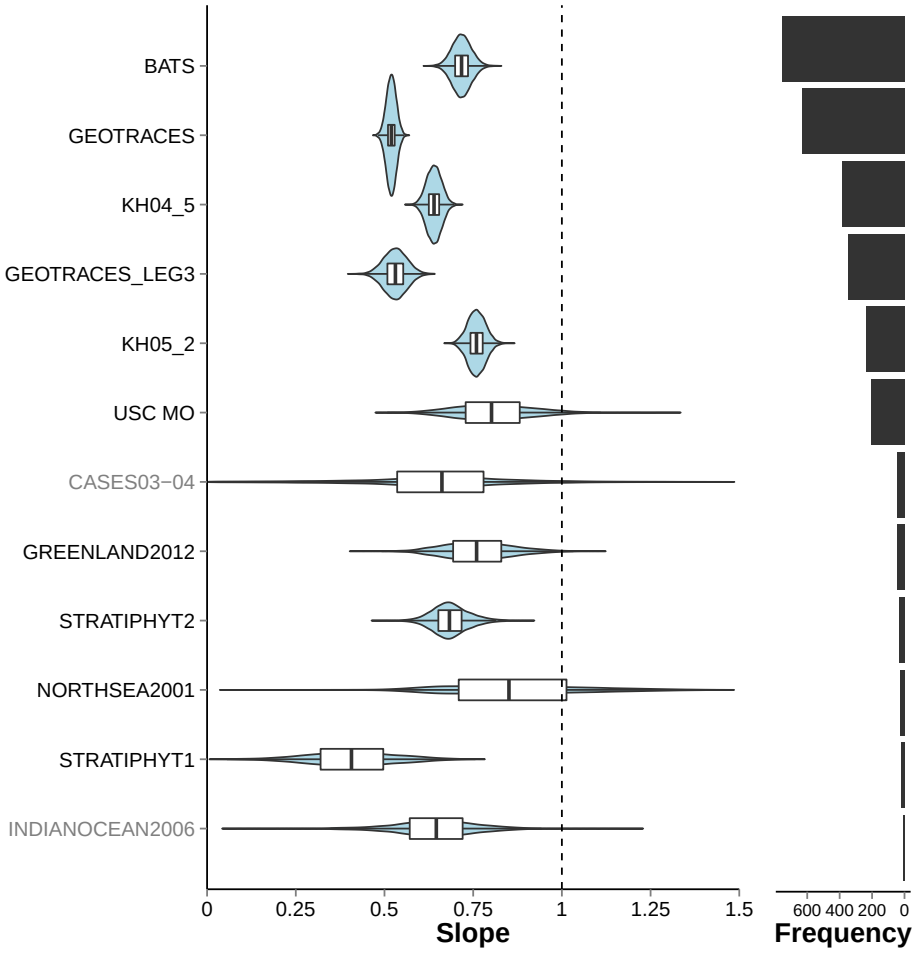


Figure 2.11: **Study-specific 95% confidence intervals of power-law exponents for relationships between virus and microbial cell abundance from samples taken at depths greater than 100m.** The confidence intervals are plotting using "violin" plots including the median (center black line), 75% distribution (white bars) and 95% distribution (black line), with the distribution overlaid (blue shaded area). The number of points included as part of each study is displayed on the right-most bar plots. Study labels in black indicate those studies whose linear regression had a p-value less than .05/12 while labels in gray indicate a p-value above this threshold.

CHAPTER 3

THE VARIABILITY OF VIRUS-TO-MICROBE RATIOS IN NATURE

3.1 Abstract

Marine viruses and microbes are critical players in global biogeochemistry yet their densities are known to vary greatly around the world. The numerical relationship between oceanic viruses and their microbial hosts has been described as sub-linear, meaning that ocean viruses are found not only at greater densities than ocean microbes, but that viral densities relative to microbial densities decrease with increasing microbial density. However, the sub-linear description of this relationship fails to fully describe the variability found in the ratio of virus density to microbe density (VMR) values. Here the variability of VMR values from 5,508 ocean water samples across 22 studies is examined and is found to increase with increasing microbial density. In contrast, the variability across studies is nearly equal. The near-constant VMR variability across studies is counter to expectations as studies differ from one another by geographic location, time of year, and the depths at which samples were taken. Nevertheless environmental covariates do not covary strongly with VMR suggesting VMR variability is not well explained by environmental factors alone. The chapter concludes with a discussion of possible interpretations for this signal.

3.2 Introducton

Marine microbes and their viruses play a central role in regulating ocean ecological and biogeochemical processes [7, 86, 18, 9]. With an estimated 1.3×10^{29} microbes [87] and 4×10^{30} viruses [88] in the world's oceans, both are highly abundant. While microbes and viruses are measured in micrometers, recent research shows that roughly "one quarter

of the anthropogenic carbon dioxide emitted on earth in the last 20 years was taken up by the ocean" [89]. This carbon sink is thought to be driven in part by the 'viral shunt',[90], the process whereby viruses infect and lyse host microbes (up to 40% of ocean microbes, daily), spilling dissolved organic matter (DOM) into the ocean where it is recycled and serves as the basis for ocean food webs and the remainder falls to the ocean floor. The interplay between ocean microbes and their viruses with regard to their impact on ocean biogeochemistry is therefore an important endeavor worthy of our understanding.

The relationship between marine microbes and their viruses has been examined and described by the the virus-to-microbe ratio (VMR), a measure which describes the per-microbe density of viruses thus acting as a quantitative measure of the relationship between viruses and their microbial hosts [91] at a given location and time. Whereas high VMR values are thought to occur when viruses are particularly productive relative to their hosts, low VMR values are thought to result from high rates of viral inactivation by microbes, losses of microbes in the environment, and/or widespread host resistance to viral infection[12].

Indeed, several papers have sought to quantify VMR values empirically[92, 93], through examinations of collations of data from worldwide sites [94, 11, 12] ultimately upending the popular notion that virus densities are pegged to microbial densities, or even relate via a fixed ratio. Specifically, Wigington et al. [94] conclude that the much [ab]used 10:1 ratio used to describe marine VMR values are in fact between 2.6 and 160 in the near-surface (≤ 100m), VMR values are between 4 and 74 in the subsurface (> 100m), and virus densities are sub-linearly related to microbe densities at both near- and sub-surface sites.

This analysis provides an update to the power-law models proposed by Wigington et al. by taking the models to the next logical conclusion which is to examine VMR variability across microbe densities thus shedding light on the trustworthiness of VMR values predicted by power-law models. Specifically, (i) VMR *variance* is found to increase with increasing sample microbial density when all data is examined, (ii) VMR variability is roughly constant across microbial densities when examined by depth class (i.e. near-

surface ($\leq$100m) and subsurface ($>$ 100m)), and (iii) VMR variance which increases with microbe density can be accounted for by the differences in mean VMR across studies.

3.3 Results

3.3.1 Virus density variability is explained by a variable-variance model

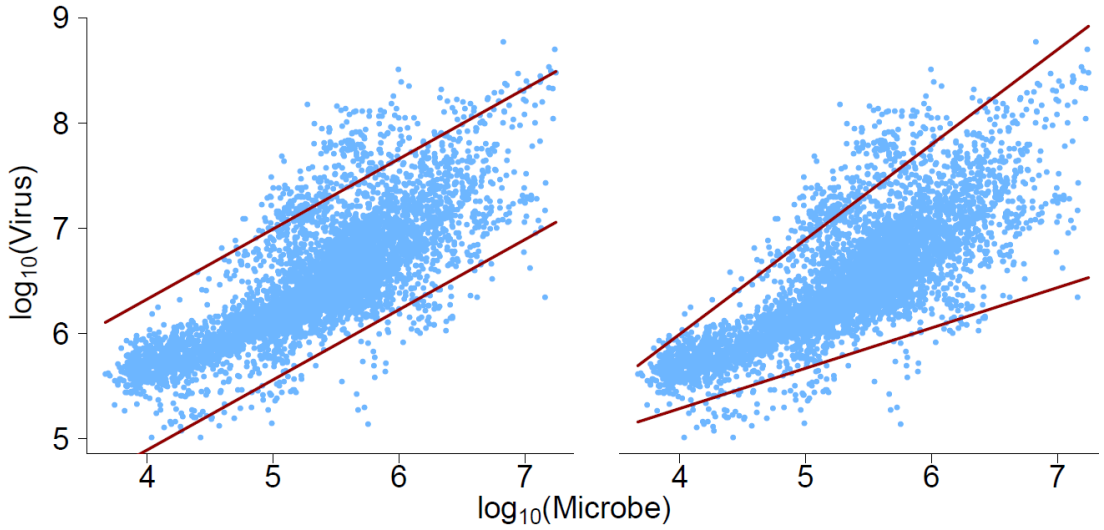


Figure 3.1: **Virus density variability is best explained by a variable-variance model.** The fit of two models for viruses density variability are contrasted whereby one model assumes virus density is constant across microbial densities while the other model allows for non-constant virus densities. The maroon lines in both panels indicates the boundary inside which 95% of all data is expected to be found, assuming normally distributed residuals.

Being that the focus of this analysis is on the relationship between viral density variability and microbial density, all of the data was examined together, regardless of environment to gain a global perspective of VMR variability. The fits of the constant-variance and variable-variance models were contrasted by AIC for all 5,508 data points contained in the dataset to identify the model which best fits the data. The constant variance model had an AIC value of 4313.368 and the variable-variance model had an AIC value of 3863.771. Of the two AIC values, the variable-variance model's is smaller, thus indicating that the variable-variance model is better described by the data and is therefore the "preferred model". Figure 3.1 shows the results of this fitting exercise in two panels, the left showing

the constant-variance model and the right panel, the variable-variance model. Both panels show microbial density (particles/ml) on the x-axis and viral density on the y-axis (particles/ml). The two red lines in either panel indicate the lower and upper boundaries in which 95% of the samples are expected to be found in microbe-density-virus-density space. Visually, it is evident by pairs of red lines that the left panel indicates the constant-variance model as the red lines are parallel to one another indicating that virus densities are expected not to vary across microbe densities.

3.3.2 Two mechanisms support non-constant VMR variability across microbial densities

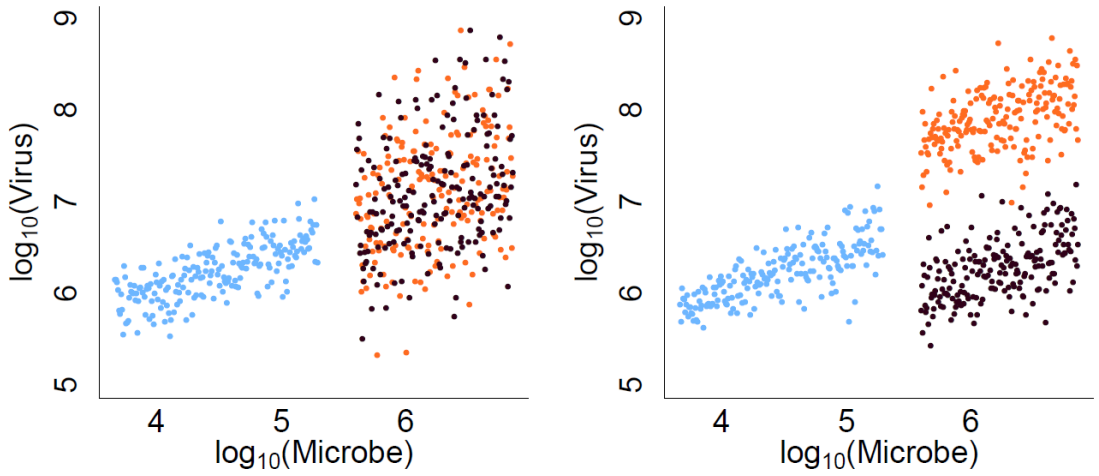


Figure 3.2: **Two mechanisms support non-constant VMR variability across microbial densities.** 1,000 log-transformed synthetic VMR values are shown in both panels. (Left) Both orange and brown studies have equal average viral densities with large variances. (Right) The orange study and the brown study have equal, smaller variances than on the left yet demonstrably different average virus densities. Both mechanisms support a the observed macro-ecological trend that a variable variance model is a better fit to the collection of data than a constant variance model.

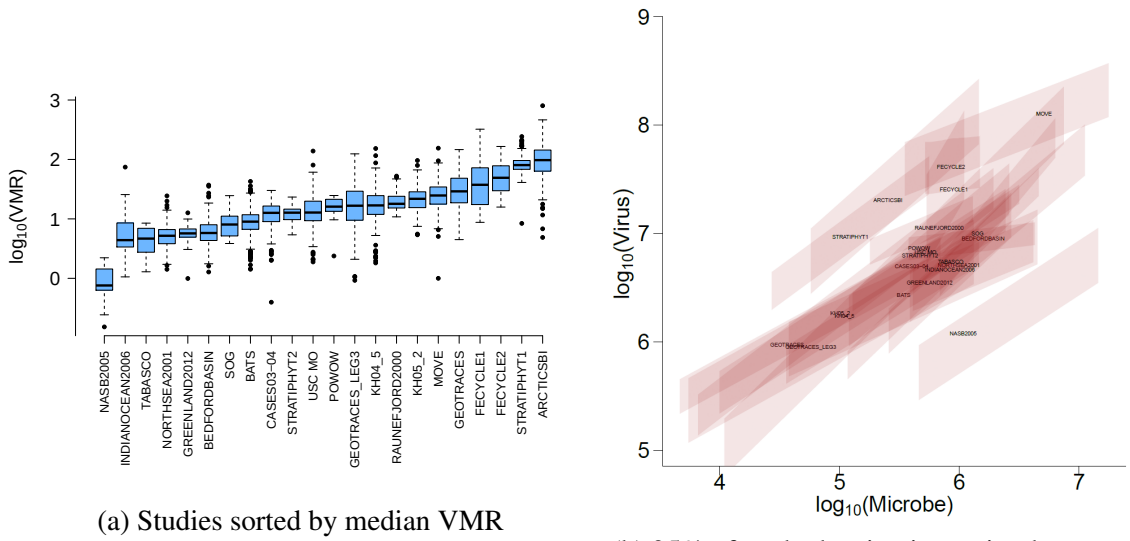
Interested to understand the mechanism by which the data might show support for variable-variance in the model, figure 3.1 shows synthetic data to visualize two causes which can create an increasing variance feature. The left panel of figure 3.1 shows three point clouds of hypothetical data which act as ”studies” and are shown in different col-

ors. The blue point cloud comes from a "study" which has evidently sampled low microbe densities (particles per ml) and low virus densities (particles per ml) and is therefore the furthest point cloud on the left of all three point clouds. The orange point cloud and the brown point cloud represent data from studies which have the same average microbial density, having sampled high microbe density sites, therefore these point clouds sit on the high end of the microbial density axis. Further, these point clouds have nearly equal mean virus densities and nearly equal virus density variances. Thus the brown and orange point clouds occupy overlapping microbe-virus density spaces but because they have much greater virus density variances than the blue point cloud, taken together the studies are best fit by a variable-variance model across the entire dataset in microbe-density-virus-density space.

The right panel of figure 3.1 also shows three hypothetical studies however where the blue point clouds are the same across panels according to microbe and virus means and variances, the high-microbe orange and brown point clouds occupy different places in the microbe-virus space. The high-microbe density point clouds in orange and brown share a common variance but very different mean viral densities: the high microbe density, high virus density study in orange sits at a higher virus density position than the high microbe density, low virus density point cloud in brown. Thus the brown and orange point clouds occupy different microbe-virus density spaces therefore when taken together with the blue point cloud, the collective studies are again best fit by a variable-variance model.

3.3.3 Median VMR values vary yet VMR variability is similar across studies

The 5,508 data points show a variable-variance model best fits the data. Figure 3.1 proposes two mechanisms by which study effects could be the driver by which the variable-variance model best fits the data. Studies were summarized by boxplots of the log (base 10) of VMR values and are shown in figure 4.8 in order from left to right showing the study with the greatest median VMR value to the study with the smallest median VMR value. The downward slope in median VMR values shows that the studies differ by over two orders



(a) Studies sorted by median VMR

(b) 95% of study data in virus-microbe space

Figure 3.3: **Median VMR values vary yet VMR is similar across studies** Sorted by median VMR value, studies have near-constant variances however the median VMR values span roughly 2 orders of magnitude across studies (left). Each study is shown with a red polygon which follows the convention that 95% of the study's data is expected to fall within the upper and lower bounds of the polygon(right).

of magnitude in terms of median VMR. Were studies to possess a similar median VMR this figure would have shown the boxplots at the same vertical position forming a nearly horizontal line of study boxplots. Likewise the variance of VMRs across studies are similar to one another as the size of the interquartile ranges (in blue) visually do not vary greatly. In fact, the distribution of variances across studies are shown in figure 3.4 highlights the narrow distribution of variances, supporting the conclusion that studies tend to have very similar variances. The only study whose variance sits outside of ± 1.96 standard deviations from the mean is the FeCycle1 study whose variance is shown at the vertical red line.

When viewed as polygons which cover 95% of the data for each study in the virus-density-microbe-density space, the similarity of the variances across studies are evident. The right panel of figure 4.8 shows each study as a red polygons where 95% of the data for each study is contained within the polygon. Viewing the data this way gives preliminary insight into the mechanism driving the fit of the variable-variance model: studies have similar variances indeed but sit in very different places according to virus density and microbe

density values. While most of the studies are concentrated in the middle of both microbe density and virus density space, the low-microbe density studies tend to have very similar median virus density value which places each study on top of one another. Yet at high microbe densities studies have similar virus-density variances however the studies are much more spread apart as can been seen by the positions of MOVE and NASB2005 which sit on opposite ends of the virus density space.

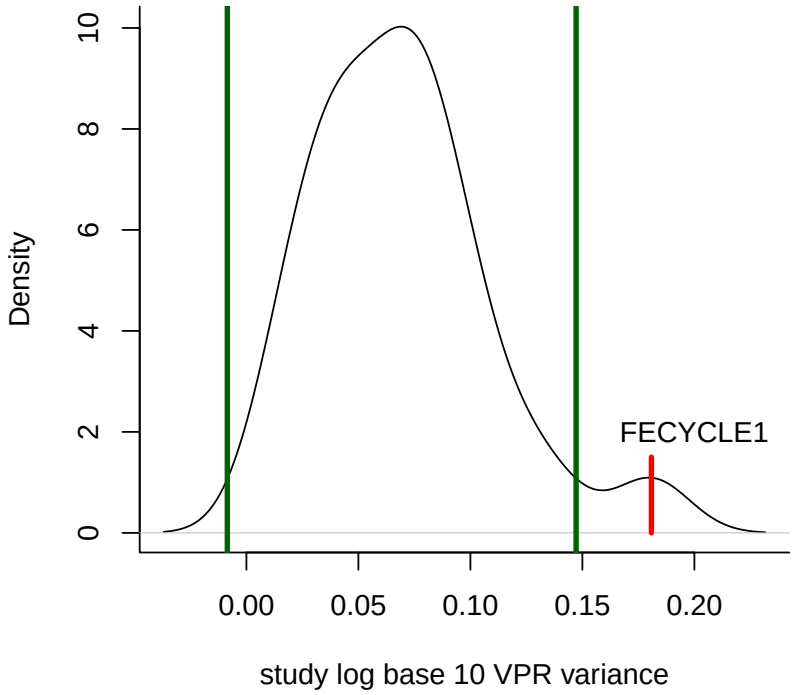


Figure 3.4: **Distribution of study variances show a small range of variances**. Study virus-to-microbe ratio variances when log-transformed, thus comparing apples to apples in terms of the magnitude of variances across virus to microbe ratios, show a tightly and approximately normally distributed set of values. The outlier among the group is FECYCLE1 which is shown to be the cause of the hump on the extreme positive end of the distribution above the 97.5 percentile cutoff.

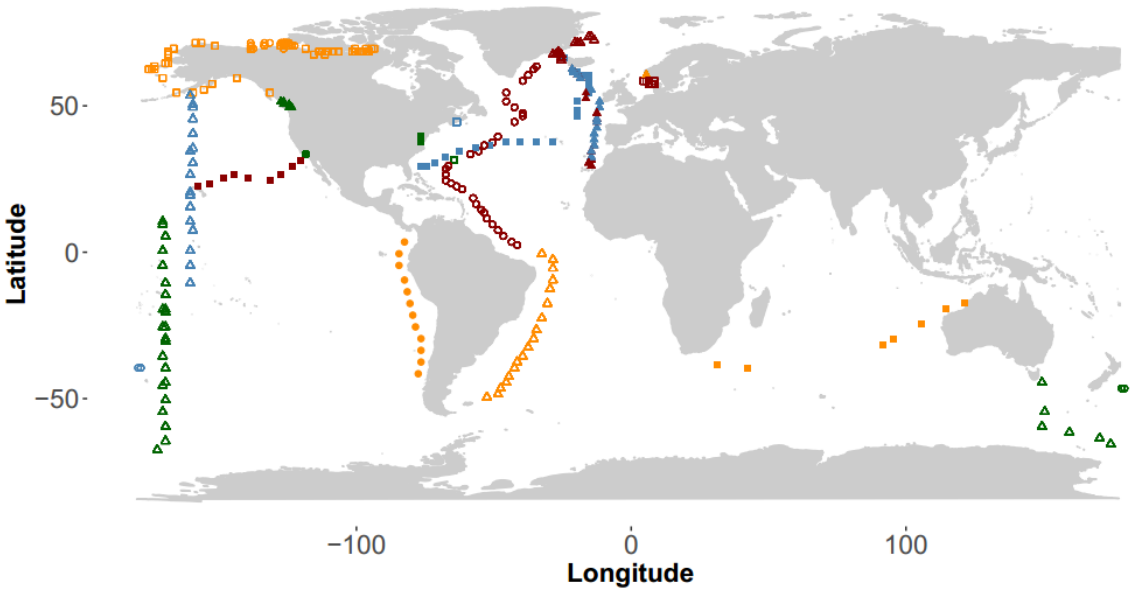


Figure 3.5: **Global distribution of samples by study hints at study effects**. Each point on the map indicates the location from which at least one sample was taken. Here 22 studies are shown, indicated by differing the shape and color of the point. The number of observations for each study are provided in Supplementary material.

Table 3.1: **Virus and microbial abundance data from 25 different marine virus abundance studies from 11 different lab groups.** A total of 5,508 data points were aggregated. The data collection dates range primarily from 2000 to 2011. Data comes from both coastal and non-coastal and both the northern and southern hemispheres, collected predominately during the summer months, with the notable exceptions of long-term coastal monthly monitoring sites (USC MO, BATS, Chesapeake Bay).

Study Name	Study Type	Location	Regime
NORTHSEA2001	Spatial	North Sea	Coastal
RAUNEFJORD2000	Temporal	North Sea	Coastal
BATS	Temporal	Sargasso Sea	nonCoastal
STRATIPHYT1	Spatial	N-Atlantic Transect	nonCoastal
STRATIPHYT2	Spatial	N-Atlantic Transect	nonCoastal
USC MO	Temporal	Santa Barbara Channel	nonCoastal
GEOTRACES	Spatial	Atlantic Transect	nonCoastal
GEOTRACES_LEG3	Spatial	Atlantic Transect	nonCoastal
BEDFORDBASIN	Temporal	North Atlantic Ocean	Coastal
GREENLAND 2012	Spatial	Greenland Sea	nonCoastal
INDIANOCEAN2006	Spatial	Indian Ocean	nonCoastal
KH04-5	Spatial	Southern Pacific Ocean	nonCoastal
KH05-2	Spatial	Northern Pacific Ocean	nonCoastal
CASES03-04	Spatial	Arctic Ocean	nonCoastal
SOG	Temporal	Pacific Ocean - Strait of Georgia	Coastal
ARCTICSBI	Spatial	Gulf of Alaska	Coastal
FECYCLE1	Spatial	South Pacific Ocean	nonCoastal
FECYCLE2	Spatial	South Pacific Ocean	nonCoastal
NASB2005	Spatial	North Atlantic Ocean	nonCoastal
POWOW	Spatial	Pacific Ocean	nonCoastal
TABASCO	Spatial	South Pacific Ocean	nonCoastal
MOVE	Temporal	Atlantic - Chesapeake	Coastal

3.3.4 Global distribution of samples by study hints at study effects

Wigington et al. show that marine virus density and marine microbe density samples in this aggregated data set were taken from around the globe therefore clues to the differences in studies shown in figure 4.8 were first examined by mapping the studies geographically, giving each study a unique combination of color and shape to set it apart from the 21 other studies in the dataset. Figure 3.5 shows a point on a map for the geographic location where a sample was taken. From this figure it is clear that samples tend to be tightly clustered according to the study which collected the sample. In total, 22 different studies are shown on the map spanning classifications beyond just near-surface ($\leq$ 100m) and sub-surface ($>$ 100m), as in the analysis by Wigington et al. This map shows the extent to which studies differ in terms of sampling location, notably Polar vs. equatorial and Pacific Ocean vs. Atlantic Ocean vs. Indian Ocean vs. Arctic. Table 3.1 (adapted from Wigington et al.) specifies the environment from which each study sampled according to location, study type, and regime highlighting the differences across studies. 27% of studies (6/22) were temporal in design and 27% (6/22) of studies took samples primarily from coastal waters (note: temporal studies are predominately but not strictly coastal studies). Putting together the differences in studies sampling site and study position in virus-microbe space, panel 4.8b of figure 4.8 shows studies NASB2005 and MOVE to be on opposite side of the virus-density space but curiously both studies come from the Atlantic ocean. Furthermore, the NASB2005 data comes from samples from the northern Atlantic Ocean yet samples taken as a part oft he ArcticSBI study come from near the Arctic Ocean yet again, the studies sit on opposite side of the virus-density space.

3.3.5 Study-specific intercepts improve variable-variance model fit

The results of modifying the variable-variance model to create two study-level models, thus adding 22 parameters to each model, are shown in the bottom left and bottom right panels of figure 3.6. All four models (one constant variance and three variable variance) are

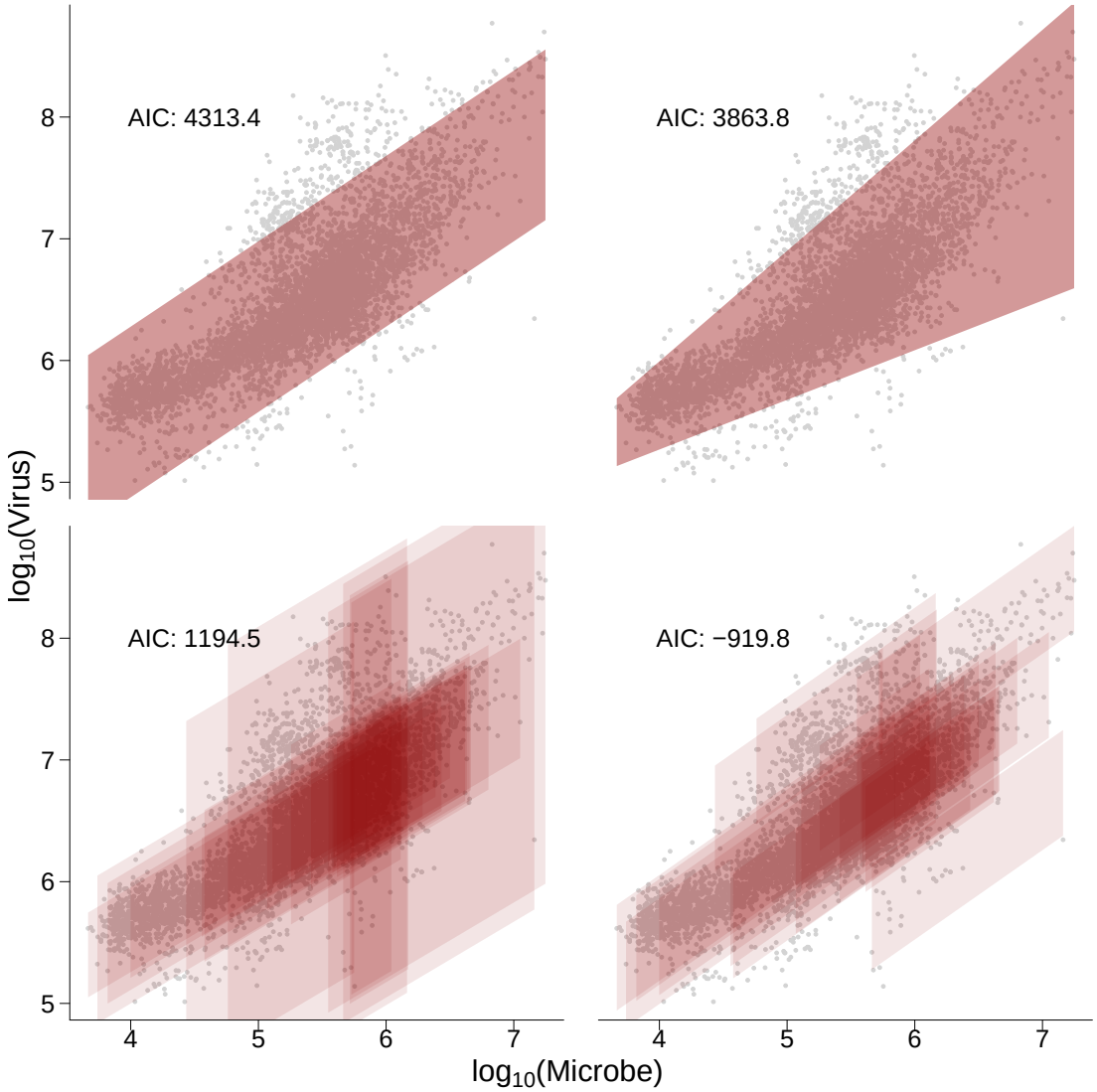


Figure 3.6: **Study-specific intercepts improve variable-variance model fit** The constant-variance model (top left) and variable-variance model (top right) were fit to viral density across the range of microbial densities. The variable-variance model was updated in two ways: first, to allow studies to share a common intercept with individual, study-specific variances (bottom left) and second, to allow for a variance value shared across studies and individual, study-specific intercepts for each study (bottom right).

contrasted by AIC values. The top left model is the macro-level constant-variance model, the top right is the macro-level variable-variance model, the bottom left is the common-study-intercept variable-variance model, and the bottom right is the study-specific-intercept variable-variance model. As expected the models with additional study-level parameters have lower AIC values than either of the macro-level models and as expected, the model which allows each study to have its own intercept and share a common variance fits the data best from among the four models (AIC = -919.8). The second best model is the one which allows each study to have its own variance and share a common slope and intercept (AIC = 1194.5).

3.4 Discussion

The relationship between viruses and their marine hosts is increasingly studied to better model the effect of viruses on global biogeochemical cycling. The description of the numerical relationship between viruses and microbes to be a 10:1 ratio has been shown to be rarely accurate [94] and in this dataset is observed to occur in less than 5% of samples. Quantifying the variability in this relationship is expected to shed light on the degree to which predicted virus densities, by knowledge of microbial densities, are reliable. Here we show that across all microbial densities, not only do virus densities increase but so too does viral density variance. Two simple models show that an increase in virus density with microbial density means that predictions of virus density from microbe density are more accurate when microbe densities are small, somewhere on the order of 10^4 microbes per mL and predictions of virus density from microbe density are less accurate as microbe densities increase.

We propose two mechanism to explain the decrease in reliability with increasing microbial density. The first supposes that studies experience that which is observed across all studies, i.e. studies which sample low microbial environments observe low virus density variances while studies which sample high microbial-density environments observe

roughly the same average virus density but large amounts of variation across samples. The second mechanism supposes that the trend observed across studies is the result of the effect of analyzing studies which are simply very different from one another. Thus, virus density variation within studies is small and mean virus density variation across studies is large and in fact increases with average microbial density. We show that indeed the second mechanism better explains the data and virus-to-microbe ratio variation across studies is relatively constant with only one outlier.

Considering the model which allows for increasing virus density with increasing microbe density as a result of study-effects, contrasting two variable-variance models which include study-specific mechanisms for different virus densities supports the conclusion that studies observe similar virus density variances but very different virus density medians. Notably the data in this analysis come from around the world which might suggest that the environment from which samples were taken could play a role in determining study median VMR values. Unexpectedly, studies from two polar environments sit on opposite sides of VMR space and one polar and one temperate study sit very closely in VMR space.

3.5 Conclusions

The power law model proposed by Wigington et al. [94] showed that VMR values increase sub-linearly, i.e. with increasing microbial density VMR values increase at a diminishing rate. The relationship between VMR values and microbial density is evident in supplemental figure **??** and although a power-law model sufficiently describes the relationship between points estimates of viral density and microbial density, it is also clear that increasing microbe density occurs with increasing VMR variability.

The model of VMR proposed by Wigington et al. [94] failed to consider the observed trend in VMR variability where most statistical models presume errors are uncorrelated with any predictor variables, meaning that in a linear model the residuals are normally distributed around the line of best fit and that the shape of the normal distribution around

the line of best fit is constant.

The results shown in this analysis describe VMR as increasing in variability with increasing microbe density and that the driver of this increasing variability is resulting from differences in median VMR values observed across studies. Yet, *individually* studies show VMR variability is constant across microbial densities. Both qualitative and quantitative results show limited differences in VMR variability from study to study despite studies coming from different environments around the world. The similarities in VMR variability across studies suggests further attempts must be made to uncover environmental factors which might be driving VMR variability.

Describing the variability of VMR is an improvement on previous efforts to characterize the relationship between microbes and viruses in marine environments [94]. We observe that variation in VMR (as described by σ) increases with microbial density for data from all depths (Figure 3.1), VMR variability is constant across microbial densities when examined as near-surface ($\leq$100m) and sub-surface ($>$100m) samples, studies which include samples described as high-microbe experience different variability in their relationships with virus densities, and while the environment affects samples in different geographic locations differently, the water temperature from which samples are taken impacts VMR globally.

3.6 Methods

3.6.1 Data and Computing

The data and R code used in this analysis is archived at [95] as well as being available through Github at

https://github.com/WeitzGroup/VMR_variability.

The analyses contained in the manuscript were conducted with R version 3.2.4 available at https://cran.r-project.org/. The maximum likelihood estimation methods used in this analysis employed the R package *bbmle* version 1.0.18.

3.6.2 Variance parameter identification

Two variable-variance models were created to further examine the variance of the studies. Based upon the variable-variance model, the first modified variable-variance model adds one parameter for each study so that the studies all share a common intercept and differ according to individual, study-specific variances. The second modified variable-variance model differs the original model by adding one intercept parameter for each study so that the studies all share a common variance and differ according to individual, study-specific intercepts. The dark red lines to indicate the boundary within which 95% of the data is found was calculated by assuming that 95% of the data falls within 1.96 times the size of the variance, below and above the mean. The assumption that the virus densities are normally distributed comes from the assessment shown in supplemental figure **??**.

3.6.3 Departures from linear regression

We assumed that VMR data are created by a process whereby the logarithm of microbial density determines the logarithm of virus density. This process can be described by the following regression model:

$$\mathbf{y} = \alpha + \beta\mathbf{x} + \epsilon \tag{3.1}$$

where $\mathbf{y}$ and $\mathbf{x}$ represent vectors of data for virus and microbe density, respectively and the error is normally distributed with location zero and scale equal to sigma, i.e. $\epsilon \sim N(0, \sigma)$. For a simple linear regression model, values for the intercept α and the slope β can be solved numerically by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3.2}$$

which minimize the square of the error between the predicted virus densities and the observed virus densities. Interpretations of the β value in this model hinge on the assumptions about ϵ, i.e. that ϵ is both normally distributed and that the variance σ of the normally distributed errors is constant. To understand how virus density variability varies with microbial

density, the assumption of normal and constant errors is intentionally violated such that the variance of the error σ is a linear function of x, or

$$\sigma = \gamma + \delta \mathbf{x}. \tag{3.3}$$

As mentioned above, with the observed viral density and microbial density data, the α and β values are easily identified by equation 4. Maximum likelihood methods from the *bbmle* package in R were used to identify the values of γ and δ.

To understand the contribution of studies to the global trend of increasing variance, the total viral density variance was broken down into the sum of the variances of each of the studies. This was modeled in two ways; first as the sum of 22 studies each with the same intercept and slope, differing by the variance, and second by allowing each study to have its own slope and intercept with a common variance shared by all studies. The study-specific models were fit using models adapted from the maximum likelihood methods described above (with 22 more parameters per model).

3.7 Acknowledgments

This work was supported by a grant from the Simons Foundation (SCOPE Award ID 329108 awarded to J.S.W.) and is a contribution of the Simons Collaboration on Ocean Processes and Ecology.

3.8 Supplemental figures

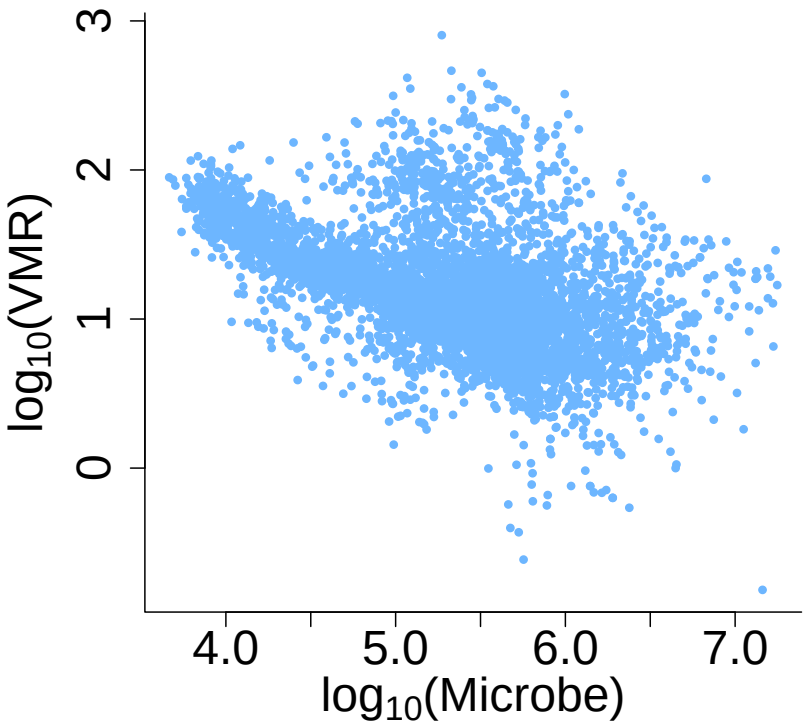


Figure 3.7: **VMR values decrease with increasing microbial density.** The sub-linearity of the relationship between microbial density and VMR is evident from the downward curve in the microbe-density-VMR space.

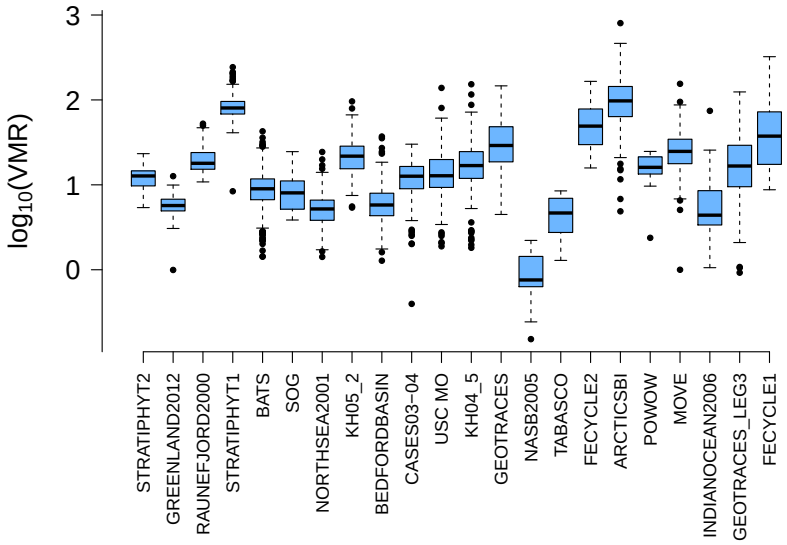


Figure 3.8: **VMR variability is similar across studies yet median VMR values vary** Sorted by VMR variance, a relationship between VMR variability and median VMR value is not apparent.

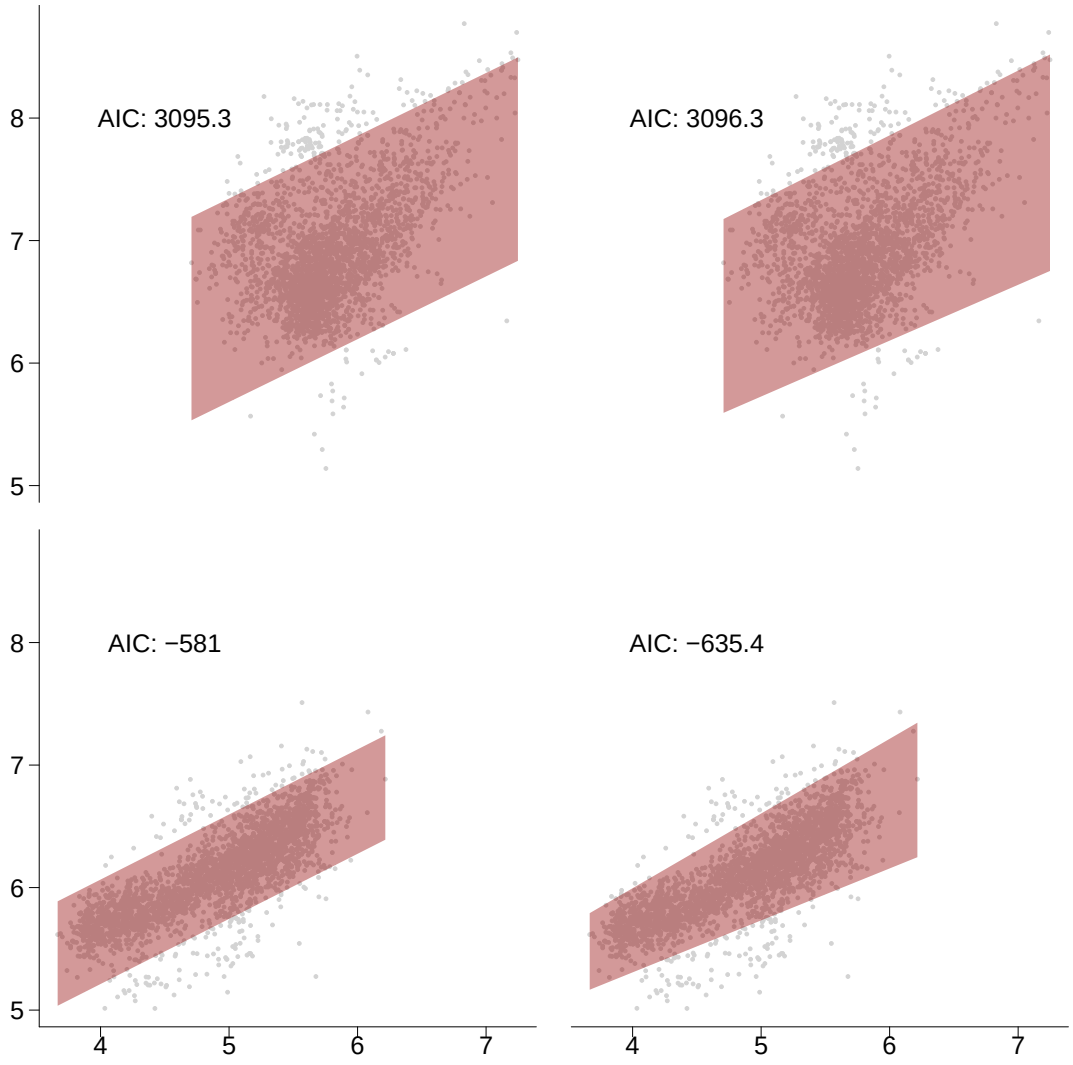


Figure 3.9: **Variable-variance models fit the data better than constant-variance models only for sub-surface.** Constant-variance and variable-variance models were fit to both near-surface and sub-surface data. By row, the model with the lower AIC value is the preferred model. The constant-variance model fits the data better in the near-surface ($\leq$ 100m) while the variable-variance model fits the data in the sub-surface ($>$ 100m) better.

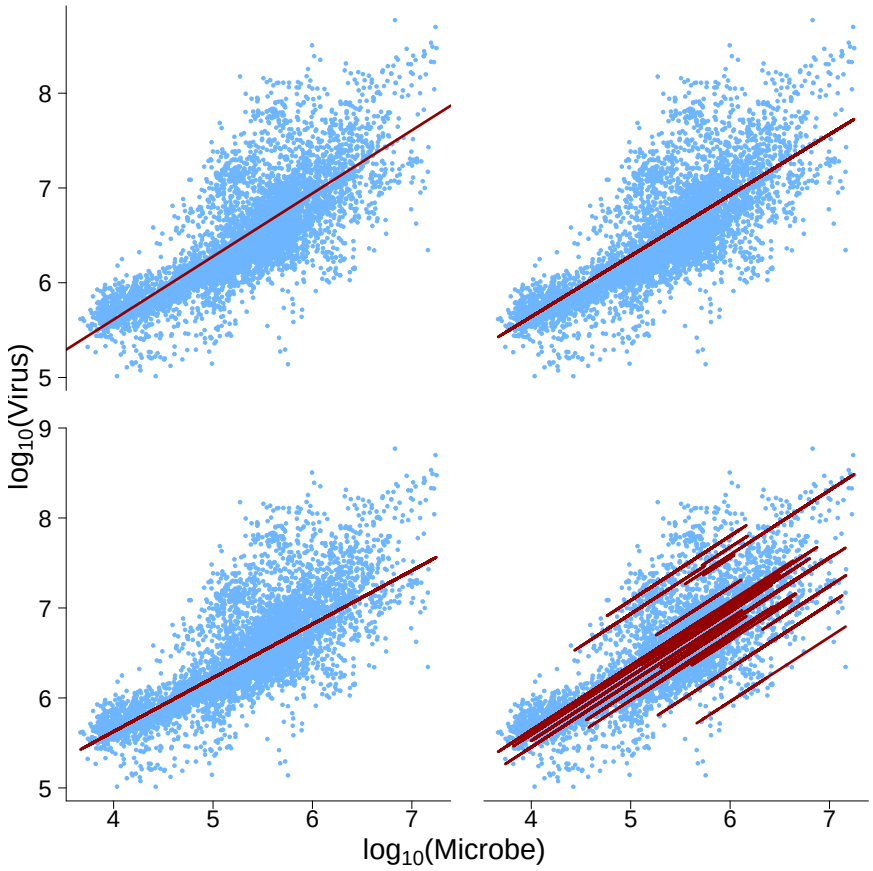


Figure 3.10: **Predicted virus densities are similar across variance models.** Constant-variance and variable-variance models were fit to all data, including the study-specific variable- and constant- variance models. R^2 was used to determine the predictive ability of each model. The constant-variance model had the lowest R^2 value (top left), the variable variance model had the second lowest R^2 value (top right), the study-specific variable-variance model had the second largest R^2 value (bottom left), and the study-specific constant-variance model had the greatest R^2 value (bottom right).

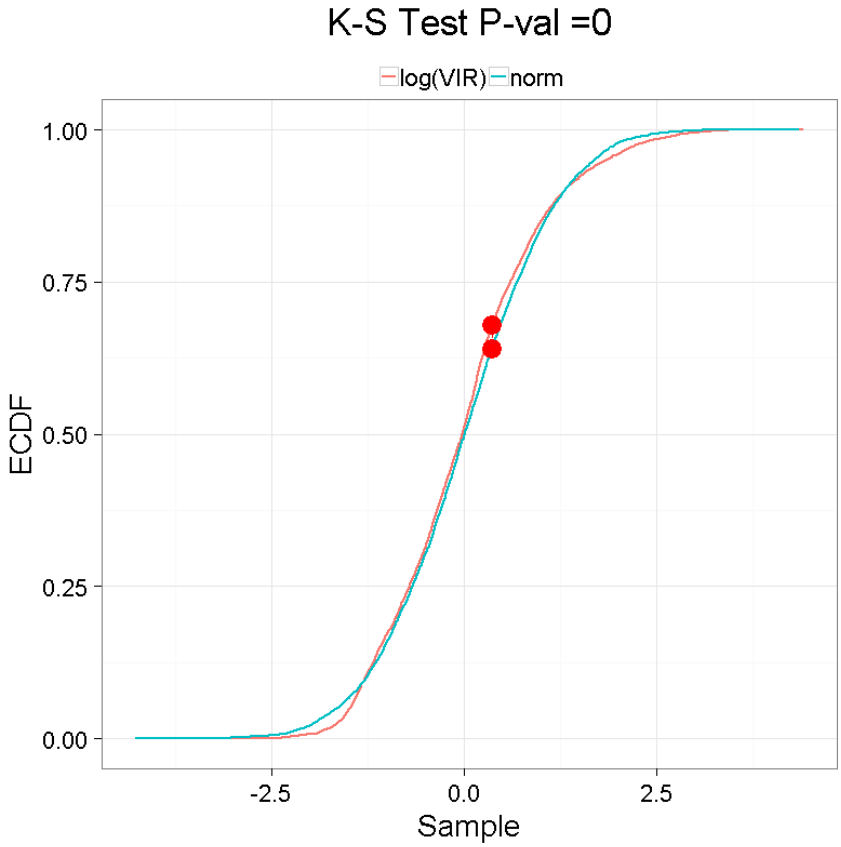


Figure 3.11: **Virus to microbe ratio values are log-normally distributed.** By the Karlmogorov-Smirnov test, taken together all data fit a normal distribution.

CHAPTER 4

ENVIRONMENTAL COVARIATES OF VIRUS-TO-MICROBE RATIOS

4.1 Abstract

Marine viruses play an important role regulating marine biogeochemical cycles. Studies show that virus densities in marine environments are not only non-linearly related to microbe densities but the variability of marine virus densities increase with increasing marine microbe densities. While the relationship between marine viruses and microbes is gaining increased attention, the role the environment plays in modulating the relationship between marine viruses their host microbes is still an active area of investigation. Research shows that nutrient availability in the ocean impacts microbial growth which in turn affects virus densities. The effect of the environment on the ratio of marine viruses to microbes was analyzed by statistical analysis of 5,508 ocean water samples which include virus density, microbe density, and environmental covariates such as salinity, temperature, longitude, latitude, depth, study type, and coastal regime. This variability of sampling locations provides a unique opportunity to use statistical methods to model and come away with insight regarding the relationship between environment and VMR. The effect of environment on VMR is not perfectly clear. No obvious linear or non-linear relationship is observed. The result of a study-effect causing samples within studies to be highly cohesive even when samples across studies are taken from the same location and depth. As a control to the potential study-effect in the data, the data was cast into Longhurst provinces to determine if biochemically similar areas best inform data cohesion. Surprisingly, analyzing the variable variance and constant variance models by Longhurst provinces instead of study showed that removing inter-study-effects may indeed be an important precursor to gaining meaningful insight into which environmental factors drive virus to microbe ratios.

4.2 Introducton

Roughly five percent of all marine biomass is estimated to come from ocean viruses yet their relatively small biomass undervalues the enormous impact they have on ocean environments, namely through microbial infection and lysis. Lysis of microbial cells in the ocean causes the organic matter within microbial cells to become available for other microorganisms to take up, causing roughly as much microbial cell death as grazers in the oceans, and sometimes more. The organic matter which is not taken up by other microbes sinks to the sea floor thus removing gigatons of organic material from the carbon, nitrogen, and phosphorous cycles. The process by which nutrients are made available to marine microbes clearly contribute to marine microbe growth but also have an impact on ocean virus levels as well[96]. This interplay between marine viruses, marine microbes, and the environments in which they thrive cuts both ways as "direct effects of changes in ocean temperature and chemistry may alter the physiological functioning, behavior, and demographic traits (e.g., productivity) of organisms, leading to shifts in the size structure, spatial range, and seasonal abundance of populations." [97] Ocean nutrients such as nitrogen and phosphorous are shown to have a significant effect on marine viral abundances such that models which use microbe abundance to predict virus abundance were improved upon markedly by the inclusion of environmental features [98]. This observation that environmental features positively correlate with virus abundance is not entirely novel as microbial abundance, chlorophyll-α abundance, and eutrophication are all known to positively correlate with virus abundance [99]. Furthermore, studies show that temperature directly affects virus inactivation rates which can occur at high temperatures while high saline environments can have a significant impact on viral community composition[98].

Here data were analyzed to examine the relationship between the environment and the ratio of viruses to microbes in marine environments (VMR). The data used in this analysis comes from samples taken around the world by different studies which differ by location

and in particular by ocean, hemisphere, gyre, season, coastal proximity, and sample depth. We begin by showing how VMR relate to environmental features and identify those environmental features which allow for the greatest discrimination between high VMR and low VMR samples. Biologically relevant environmental variables are then examined to determine if a critical set of environmental variables exists which sufficiently describes the larger set. The significance of biologically relevant variables for predicting virus density was examined at both global scales and at the study level, which led to the mapping of the studies in this data set to sections of the ocean which are thought to be biochemically similar, thus allowing comparisons of virus to microbe ratios across studies. VMR variance is examined across biochemically similar ocean segments and is seen to be qualitatively different than VMR variability across studies. The relationship between virus density variability and microbe density is then used to examine the extent to which biochemically similar sections of the ocean drive global virus and microbe density trends. Analyzing the relationship between viruses and microbes by these discrete ocean segments raises the issue that the relationships shown in studies is qualitatively different than the relationships shown in biochemically similar ocean segments. The qualitative difference in the results suggest that great care must be taken when analyzing environmental features at the global level, as latent-variable effects were identified.

4.3 Results

4.3.1 VMR samples come from a wide range of environments

Ocean water samples were taken from various locations around the world from a variety of environments including polar, tropical, coastal, open ocean, summer, winter, near-surface ($\leq$ 100m), and sub-surface ($>$ 100m) to a few. Figure 4.1 shows a set of nine scatter plots where the x-axis in each plot is a different environmental variable and the y-axis is the log (base 10) of virus to microbe ratio. While there is to be a loose relationship between temperature and VMR, there is not any one variable which has an obvious linear relationship

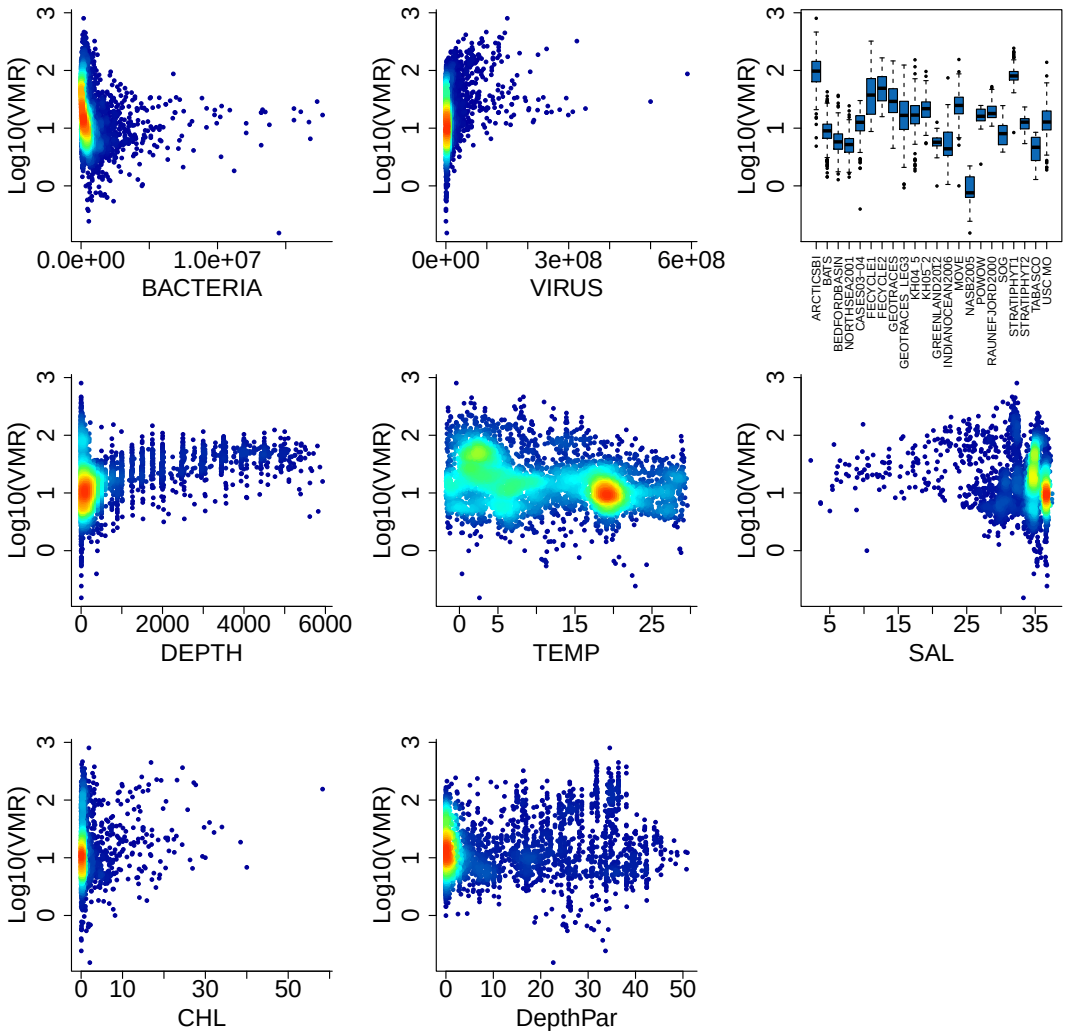


Figure 4.1: **Water samples were taken around the world resulting in a variety of sampled environments.** 5,508 marine water samples were taken from around the world yielding a range of environments including polar, tropical, coastal, non-coastal, summer, winter, near-surface ($\leq$ 100m), and sub-surface ($>$ 100m).

with VMR. In fact, scatter plots appear to provide conflicting information about what is driving the relationship between environment and VMR. Specifically, in the case the relationship between month and VMR, the largest amount of variability in VMR is observed during the summer months, primarily in July. Conversely, while one would expect summer temperatures to be warmest and therefore a scatter plot showing the relationship between temperature and VMR would show the greatest amount of VMR variability in samples taken when water temperatures are warmer, in fact the scatter plot in the first column and the second row which relates water temperature to VMR shows that samples taken in the warmest waters had less variability than samples taken at cooler temperatures. The top right panel which shows the relationship between depth and virus microbe ratio indicates that there is great variability and virus-microbe ratio and shallow depths yet surfaced and less variability with increasing depth. Likewise, this negative relationship between temperature and VMR is not entirely supported by the scatter plot relating depth to VMR as VMR decreases from the surface down to 500 meters and subsequently increases down to 3,000 meters before leveling off.

4.3.2 High and low VMR differences in environment

The relationship between viruses, microbes, and the environment was also explored by examining the environmental differences observed between high VMR samples and low VMR samples. Figure 4.2 shows the differences in the densities of the variables depth (in meters), photoactive radiation (PAR), temperature, and microbe density for high and low VMR samples. The densities of each variable for the top 10% and the bottom 10% of VMR values are colored blue and red respectively. Records within the top 10% of VMR values, so-called ”high VMR samples” had an average PAR of 23.3, an average temperature of 8.92 degrees Celsius, and an median microbe density of 225,500. Records within the bottom 10% of VMR values, so-called ”low VMR samples” had an average PAR of 14.9, an average temperature of 15.2 degree Celsius, and an average microbe density of 951,500.

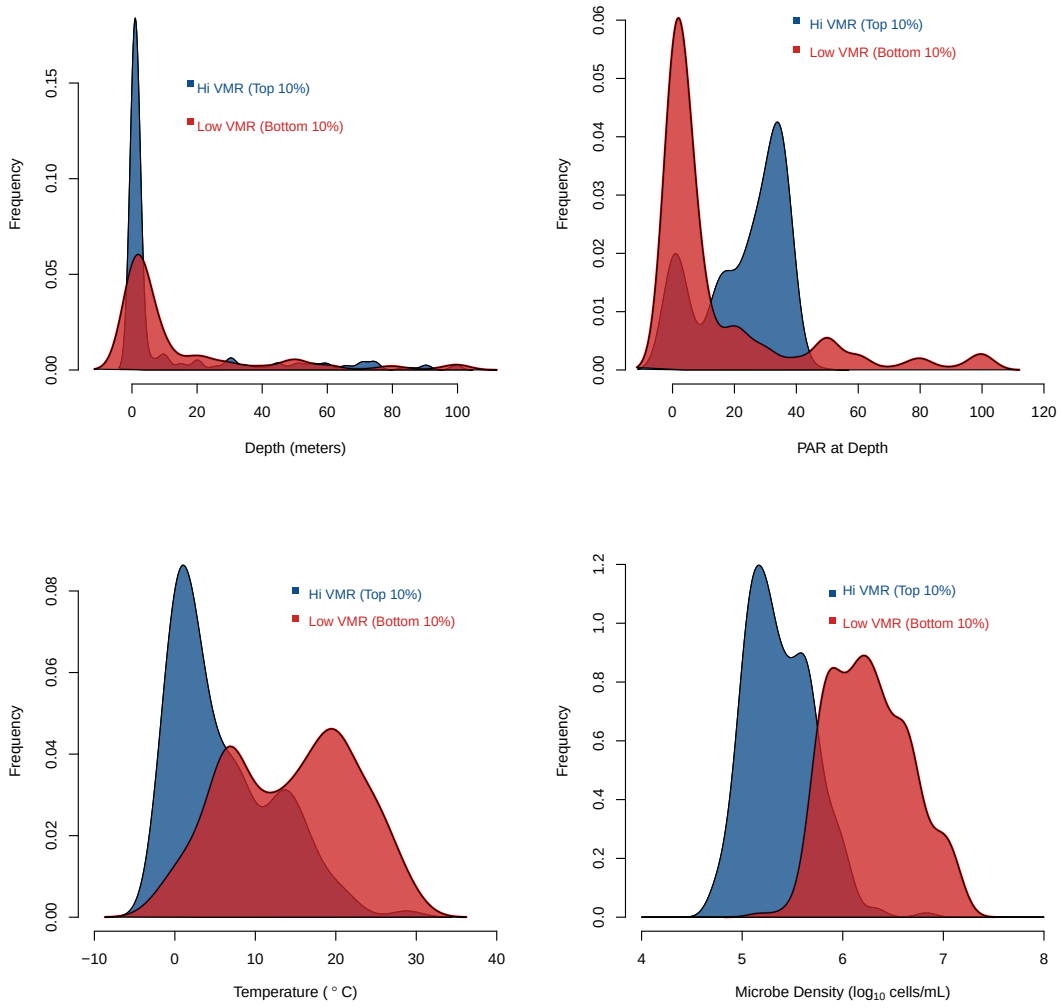


Figure 4.2: **High and low VMR differences in environment** The top 10% and bottom 10% of VMR samples are most clearly distinguished by differences in te density plots of the variables photoactive radiation, temperature, and microbe density at the time at which samples were taken.

4.3.3 PCA shows low covariation between environmental variables

Table 4.1: **PCA loadings shows low covariation between environmental variables.** Principal component analysis of environment variables taken during sampling.

Variable	PC1	PC2
Salinity	0.42	0.15
Depth	0.39	-0.15
Temperature	0.34	0.36
Mixed Layer Depth	0.31	-0.19
Surface PAR	0.14	0.62
Longitude	0.05	-0.26
Day of Year (doy)	-0.11	-0.20
Microbe density (log10)	-0.24	0.43
Latitude	-0.27	-0.24
chlorophyll-α (log10)	-0.38	0.17
Virus density (log10)	-0.39	0.17

The covariation between variables in the dataset was examined by PCA and is shown via biplot in figure 4.3. This figure shows that the first principal component explains only 33.9% of the variation in the data while the second principal component explains 14.1% of the variation in the data. Two variables which have zero covariance are shown to have perpendicular arrows such as longitude and latitude. Variables which strongly covary are shown to have overlapping (or nearly overlapping) arrows, as it is the case for the log (base 10) of chlorophyll-α and the log (base 10) of virus density. Notably, the arrow for the log of chlorophyll-α is shown to be in the opposite direction of the arrow for the variable depth, indicating that as depth values increase the log of chlorophyll-α decreases. Also, it should be noted that the arrow for the variable temperature is at an angle greater than 90 degrees to the arrow for the log (base 10) of bacteria, indicating that microbial population densities have a week negative relationship with the water temperature from which the sample was taken. Further, it is informative that there is a stronger negative relationship between the log (base 10) of microbial density and salinity than the log (base10) of microbial density and temperature. The arrows shown in the biplot for each of the environmental variables indicate that each of the environmental variables in the dataset are not well described by any

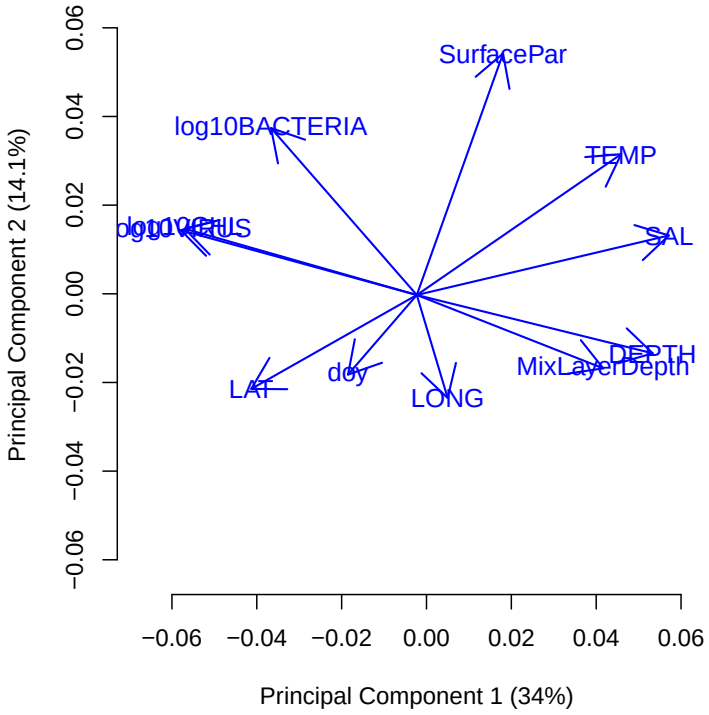


Figure 4.3: **PCA shows low covariation between environmental variables.** By principal component analysis environment variables such as longitude and latitude which are uncorrelated appear as perpendicular arrows, indicating a near-zero covariance between the two variables. Variables which have a high covariance such as chlorophyll-α (log base 10) and virus density are shown to have overlapping (or nearly overlapping) arrows.

other environmental factor and that they each contribute something unique to the dataset.

4.3.4 Studies sampled varying locations

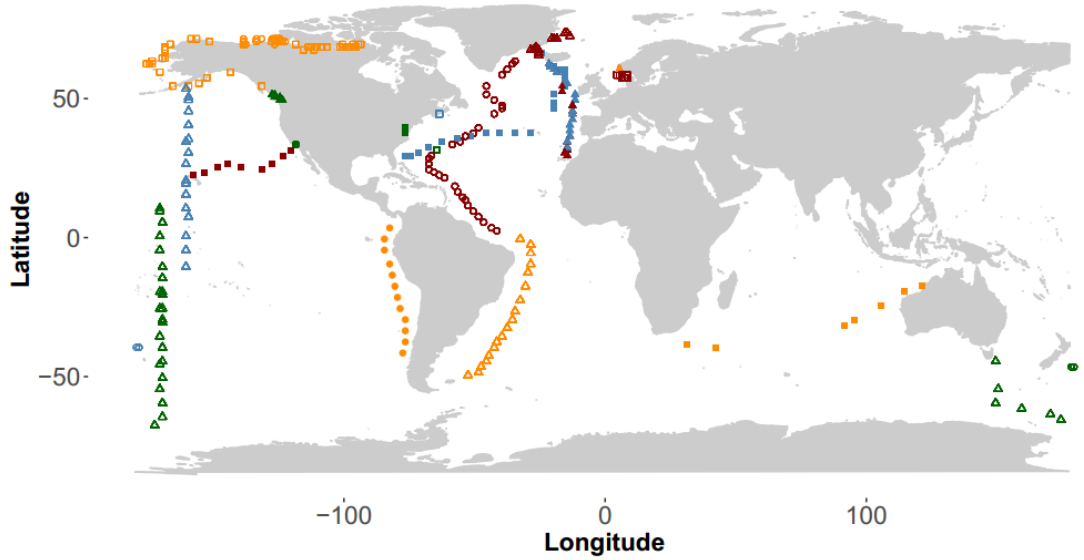


Figure 4.4: **Studies sampled from the environment in a wide array of locations but samples within studies are highly cohesive.** The collection of 5,508 records comes from a wide range of environments yet within studies, environmental variables vary little.

Figure 4.4 shows the locations where virus and microbe samples were taken. All studies in the data fall into one of two types, either time-series studies or spatial studies, and the sampling type of each study is evident from the map in figure 4.4. Time-series studies sampled the same location over time (such as USCMO, BATS, and Bedford Basin) largely without other studies sampling the same site while spatial studies such as KH04_5 and KH05_2 sampled across large areas with little to no overlapping by other studies. The location of the sampling sites for studies conducted in the northern Atlantic indicate that some areas of the ocean were sampled more frequently than others.

4.3.5 Predictive significance of environmental covariates differs across studies

Because samples were taken by different studies and therefore taken in different environments the role environmental covariates play in predicting virus density are expected to

Table 4.2: **Covariate importance differs across studies.** All data and 22 study-specific multivariable regression models comprised of eight variables were fit to the data and show the significance of the environment. Significance at the $\alpha = .05$ level is shown by an asterisk (*).

Study	Intercept	latitude	longitude	depth	temperature	salinity	chlorophyll-α	microbe Density	PAR at Depth
All	*		*		*	*	*	*	*
ARCTICSBI	*	*	*				*		*
BATS		*		*				*	
BEDFORDBASIN	*	*	*			*	*	*	*
NORTHSEA2001	*	*	*		*	*		*	
CASES03-04	*	*		*				*	
FECYCLE2									
KH04_5	*						*	*	
KH05_2		*			*				
GREENLAND2012									
MOVE							*		
NASB2005			*					*	
RAUNEFJORD2000	*		*			*		*	
SOG	*				*			*	
FECYCLE1	*		*		*	*	*	*	
GEOTRACES	*	*				*	*	*	
GEOTRACES_LEG3	*	*	*	*	*		*	*	
INDIANOCEAN2006	*					*	*		
POWOW				*			*		
STRATIPHYT1	*				*	*	*	*	*
STRATIPHYT2	*	*	*		*	*	*	*	
TABASCO				*	*	*	*		
USC MO	*	*	*			*		*	
Total		10	10	5	9	11	13	15	4

differ from study to study. The results of 23 regression models to predict virus density by eight environmental variables shows the significance of each covariate in each model in table 4.2. The variables which are below the $\alpha = .05$ significance level are shown in the table with an asterisk. *Bacteria Density, chlorophyll-α abundance,* and *salinity* are the covariates which were most frequently shown to be significant in the models, present 15, 13, and 11 times, respectively. The variables *depth* and *DepthPar*, the depth at which samples were taken and the light available at that depth, were significant in less than a quarter of the models, present five and four times, respectively.

4.3.6 Longhurst provinces were sampled unevenly by studies

To more concisely examine the role of the environment as it relates to VMR, the assertion that the ocean could be carved up into biochemically similar areas was adopted [100]. The 22 studies in this analysis sampled across 30 Longhurst provinces, often sampling the same province. Figure 4.5 shows the mapping of studies to provinces illustrating the depth of sampling for each province. This mapping shows Longhurst provinces in columns and the studies which provided the data in this analysis in rows. Each blue square indicates

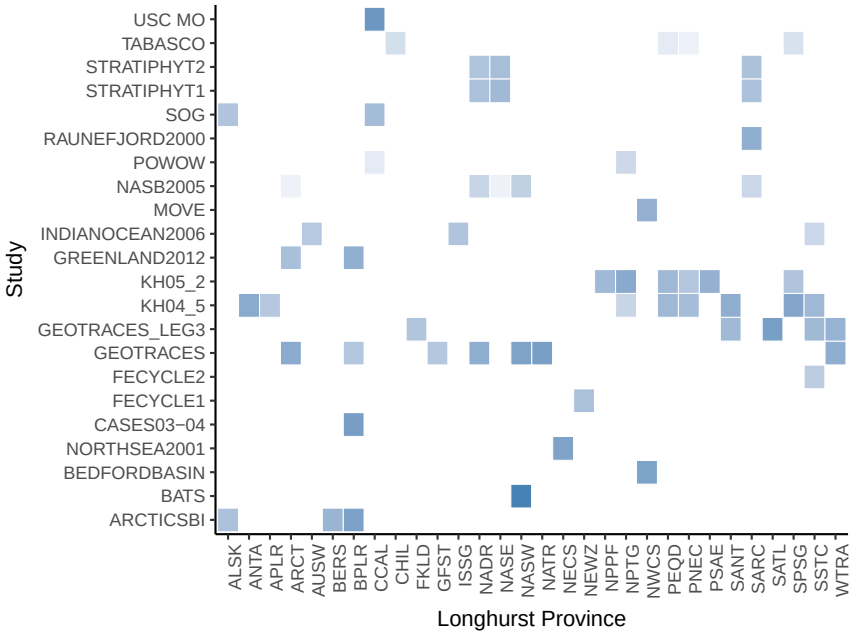


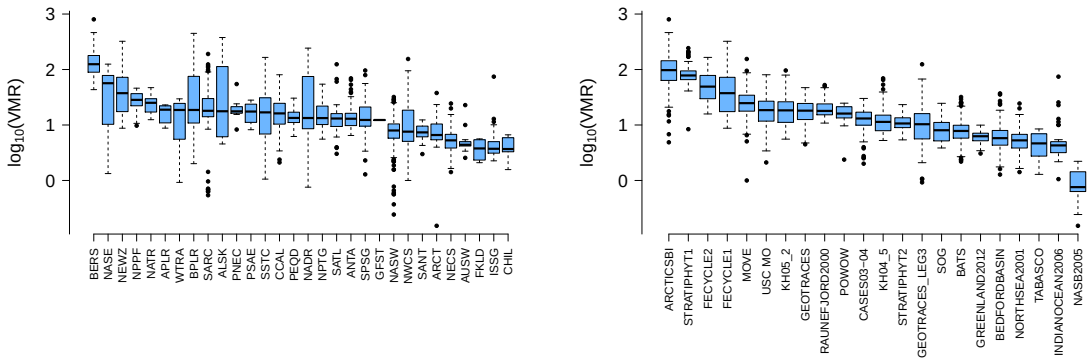
Figure 4.5: **Longhurst provinces were sampled unevenly by studies** Some studies collected samples from multiple sites, often spanning significant spatial distances. The distances spanned by studies often covered multiple Longhurst provinces. The blue squares indicate how many samples a study took within each Longhurst province.

that samples were taken by the study listed in the row within the borders of the Longhurst province listed in the column. The darkness of the squares is indicative of the number of times a province was sampled by a study with studies not sampling from a province shown in white. The distribution of Longhurst provinces sampled is skewed toward the low-end as 24 of the 54 Longhurst provinces were not sampled. Of the 30 provinces which were sampled, 13 provinces were sampled by one study and four provinces were sampled by four studies. Likewise, one study (KH04_5) took samples from eight provinces and eight studies took samples from only one province.

4.3.7 VMR variability is greater across provinces than studies

Virus to microbe ratios were aggregated according to the province from which the sample was taken. Each province's VMR values are shown by a boxplot where the bottom and top of each blue box indicates the lower and upper quartiles of the VMR values for the study

and median values are highlighted by horizontal black lines in each blue box. In order from highest median VMR value to lowest, the range of median VMR values for provinces (1.52) is less than what is observed across studies (1.87), yet by inspection the trade off is glaring: the variance in VMR values within provinces is *greater* than within studies.



(a) Provinces sorted by median VMR

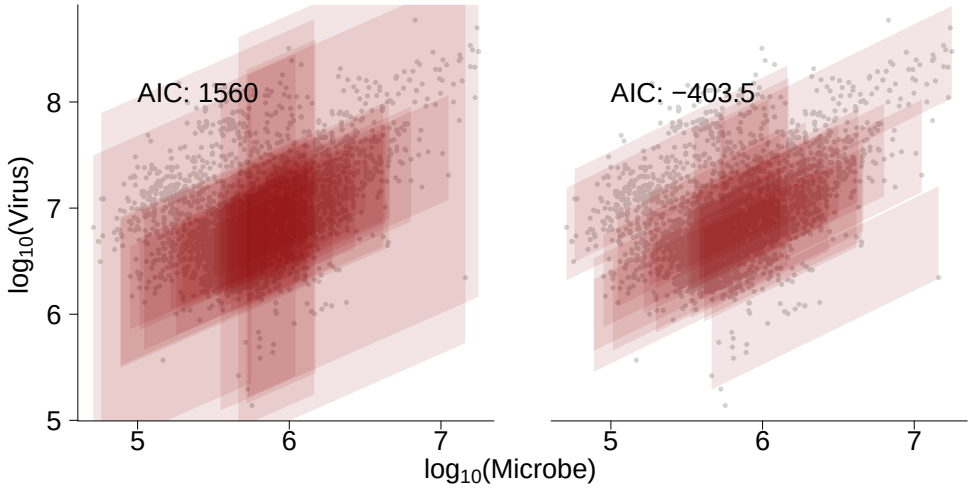
(b) Studies sorted by median VMR

Figure 4.6: **VMR variability is greater across provinces as opposed to studies** Variability in VMR is greater across provinces than across studies. The median VMR of most provinces are above a 10:1 ratio but the range in median VMR values does not span two orders of magnitude as the greatest median VMR is 2.09 in the BERS province and the lowest is .57 in the CHIL province.

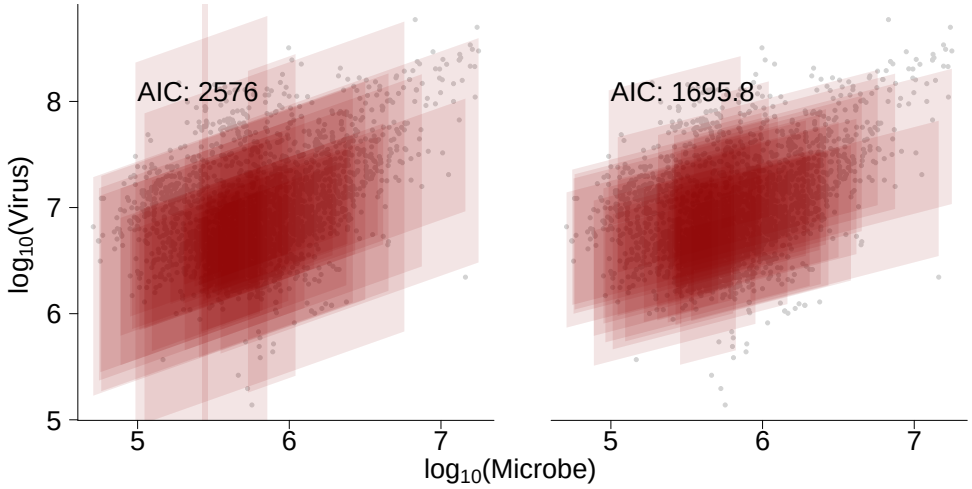
4.3.8 Variable-variance model fits provinces better than constant variance model

As Longhurst provinces describe near-surface waters ($\leq$ 100m), the constant-variance and variable-variance models proposed in chapter three were fit to the near-surface ($\leq$100m) data and show that there is a discernible difference in the fit of these two models for this set of data.

The two special cases of the variable-variance model which were analyzed in chapter three which allowed for study-specific variances and study-specific intercepts were adapted to allow for province-specific intercepts and province-specific variances and is shown as figure 4.7. It was shown previously that the study-specific intercept model fit to both near surface and subsurface data best fit the data compared to the study-specific variance model



(a) near-surface study-specific variance variable-variance model (left) and study-specific intercept variable-variance model (right)



(b) near-surface province-specific variance variable-variance model (left) and province-specific intercept variable-variance model (right)

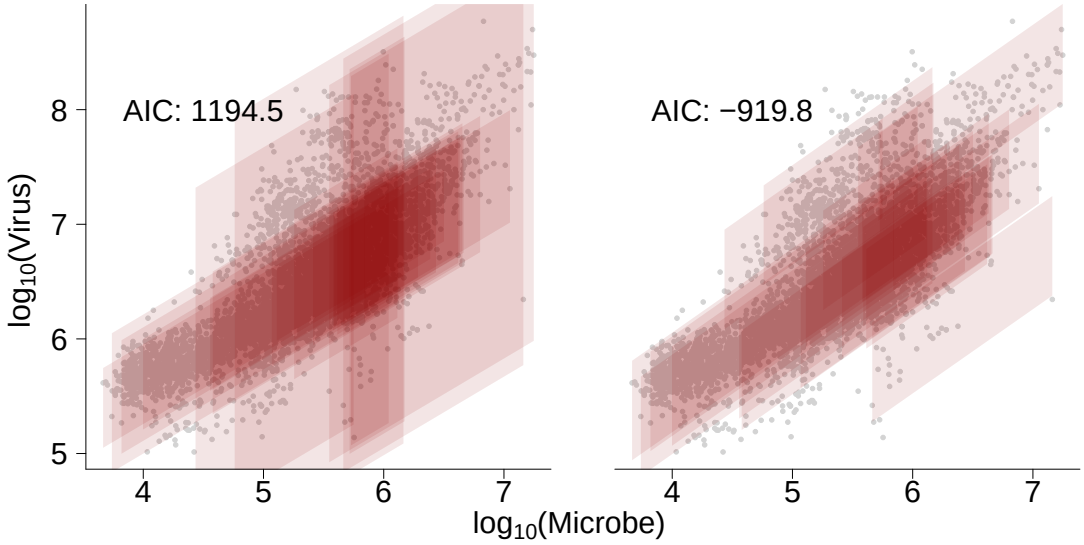
Figure 4.7: **Variability observed across studies in near-surface** Longhurst provinces are used to cluster near-surface waters around the world. Constant variance and variable variance models were fit for all near-surface samples identifying the constant-variance model as marginally better fit to the near-surface data in the top row of figure 3.9. However the study-specific intercept variable-variance model was a much better fit for near-surface data. Similarly, the province-specific intercept variable-variance model was a much better fit for near-surface data.

(figure 3.6). The findings here further support this finding as the 31 parameter models (one parameter for each of the Longhurst provinces plus a common parameter) show that the AIC value of the province-specific intercept model is less in AIC value than the province-specific variance model.

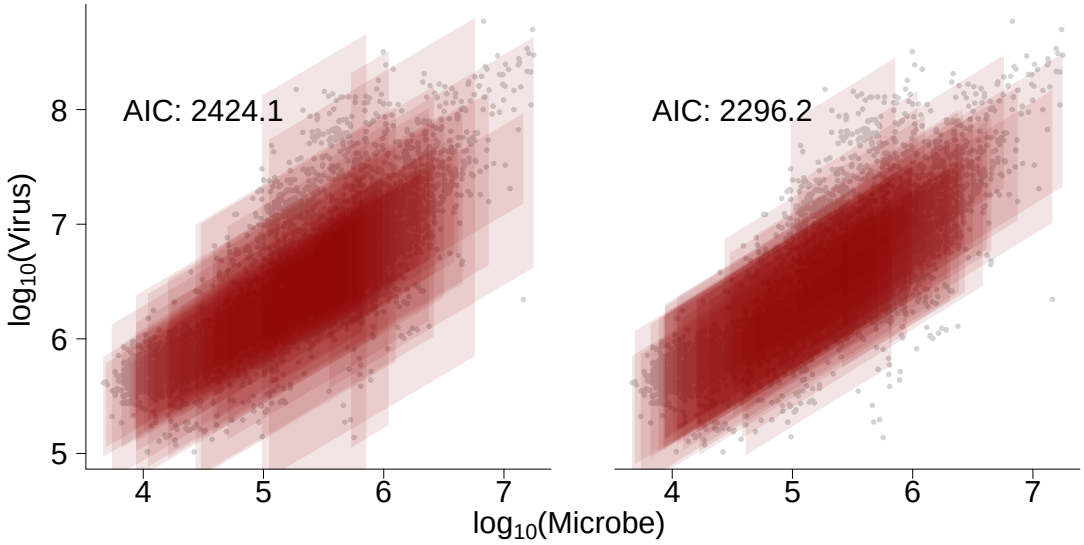
The two special cases of the variable-variance model were analyzed a second time using both near-surface and sub-surface data sampled within each province, (not just the near-surface data as in figure 4.7). Study-specific variances and study-specific intercepts were again adapted to allow for a province-specific intercept model and a province-specific variance model and is shown as figure 4.8. Again, the findings here show that the AIC value of the province-specific intercept model is less in AIC value than the province-specific variance model, as was seen for the study-specific models, fit to both near-surface data and data from all depths.

4.3.9 Virus-microbe power-law model fit to province data

A power-law model relating the base 10 log of viral density to the base 10 log of microbial density was fit to the data in each province. The virus and microbe density values for each province are shown in figure 4.9. Each panel in the figure shows where the data in each province lies in microbe-virus space in addition to a line representing the power-law model fit to that data (in blue), as well as a line which represents the power-law model fit to all of the near surface ($\leq$ 100m) data (in red), and a line which represents the 10:1 ratio (in black). The panels are in shown in the figure in order from the province with the greatest positive power-law coefficient in the top right corner to the province with the most negative power-law coefficient in the bottom left corner. 20 of the power law models fit to the data have positive coefficients while 10 have negative power-law coefficients. Upon close inspection, it is evident that the panel which shows data from the North Atlantic Drift province (NADR: column 2, row 5) shows three clusters to the data in the province. Likewise, the panels for the North Atlantic Subtropical Gyre province (NASE: column 5, row



(a) Study-specific variable-variance models



(b) Province-specific variable-variance models

Figure 4.8: **Virus density modeled by variable-variance models for all depths** Virus density is better fit by a study-specific-intercept model as opposed to a study-specific variance model. Likewise, virus density is better fit by a province-specific-intercept model as opposed to a province-specific variance model.

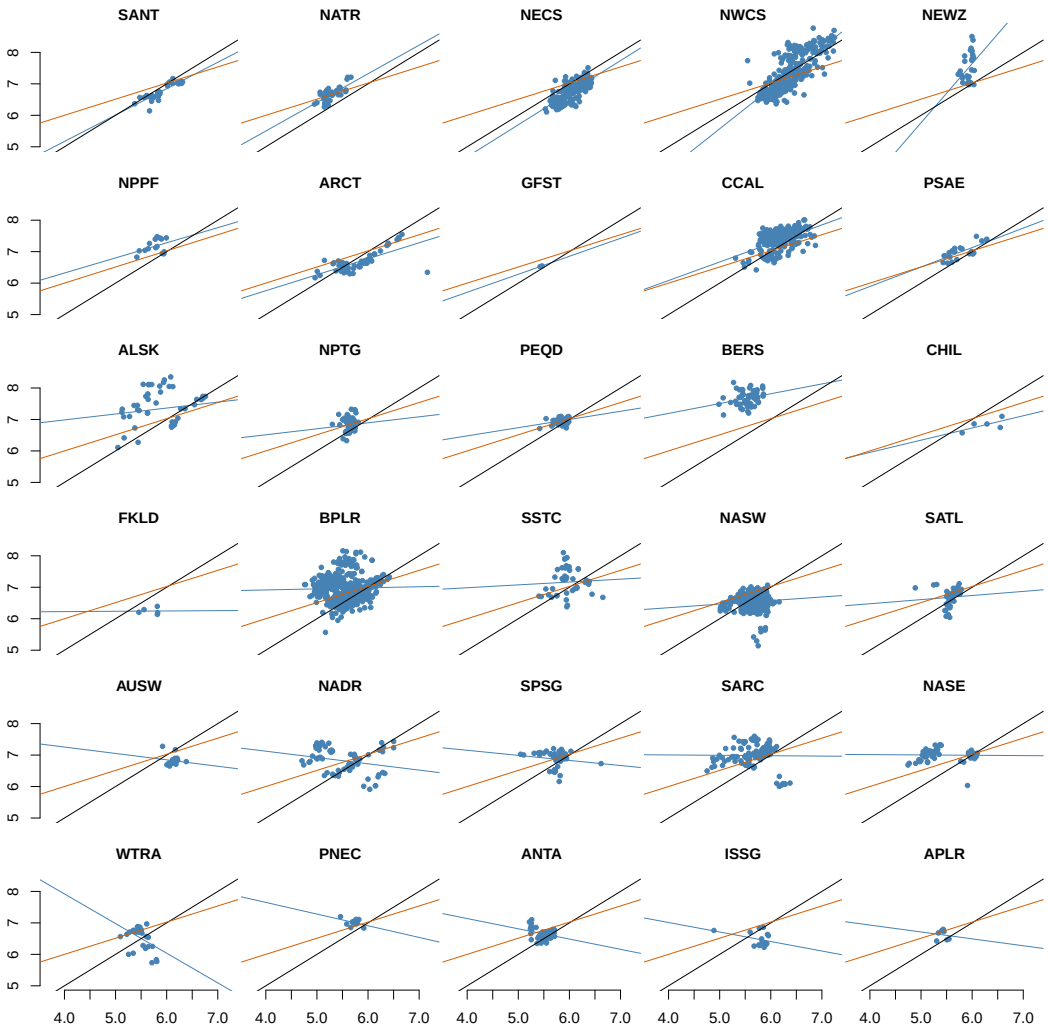


Figure 4.9: **Longhurst provinces have different virus-host relationships** The relationship between viruses and microbial hosts differ across Longhurst provinces. VMR of some provinces cluster systematically above-, at-, or below- the 10:1 VMR ratio line (in black). The orange line denotes the best-fit power law model fit to all of the data. The blue line represents a power-law model for a single study.

5), Alaska Downwelling Coastal province(ALSK: column 1, row 3), Western Tropical Atlantic province (WTRA: column 1, row 6), and the Northwest Atlantic Shelves province (NWCS: column 4, row 1) all show the data for clustered into two or more clusters within the province. Considering the mapping of studies to provinces in figure 4.5 it is not surprising that the North Atlantic Drift province (NADR), the North Atlantic Subtropical Gyre province (NASE), the Alaska Downwelling Coastal province (ALSK), the Western Tropical Atlantic province (WTRA), and the Northwest Atlantic Shelves province (NWCS) were all sampled by multiple studies.

4.3.10 Virus-microbe relationships within provinces are heavily biased

The presence of multiple studies which sampled the same province should yield nearly identical virus in microbe density values. Upon closer inspection of the virus-microbe space according to province such as the North Atlantic drift province (NADR) shown in figure 4.10 and when colored according to the study which sampled the province, it becomes clear that a latent-variable effect is overpowering the true signal of the virus-host relationship within the province.

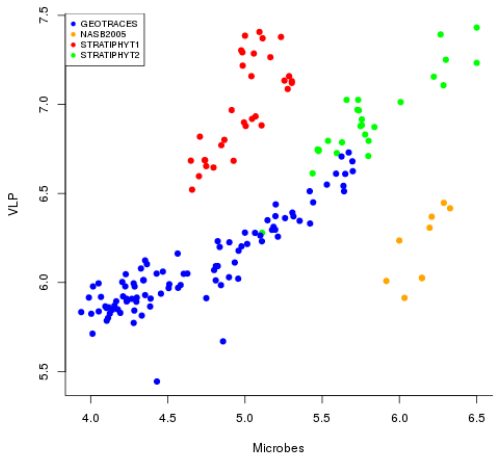


Figure 4.10: **North Atlantic Drift province (NADR) data cluster by study** The Longhurst province NADR shows the study effect first-hand in virus-microbe space as the data cluster according to the study which sampled the data.

When the data is viewed in virus density - microbe density space according to the province from which the samples were taken, it becomes clear as shown in figure 4.11 that many of the provinces which were sampled by more than one study are subject to a latent variable effect, notably the Bering Sea province, the Alaska province, the Arctic province, the Northwest Atlantic Shelves province, the Western tropical Atlantic province, and the Atlantic sub-Arctic province.

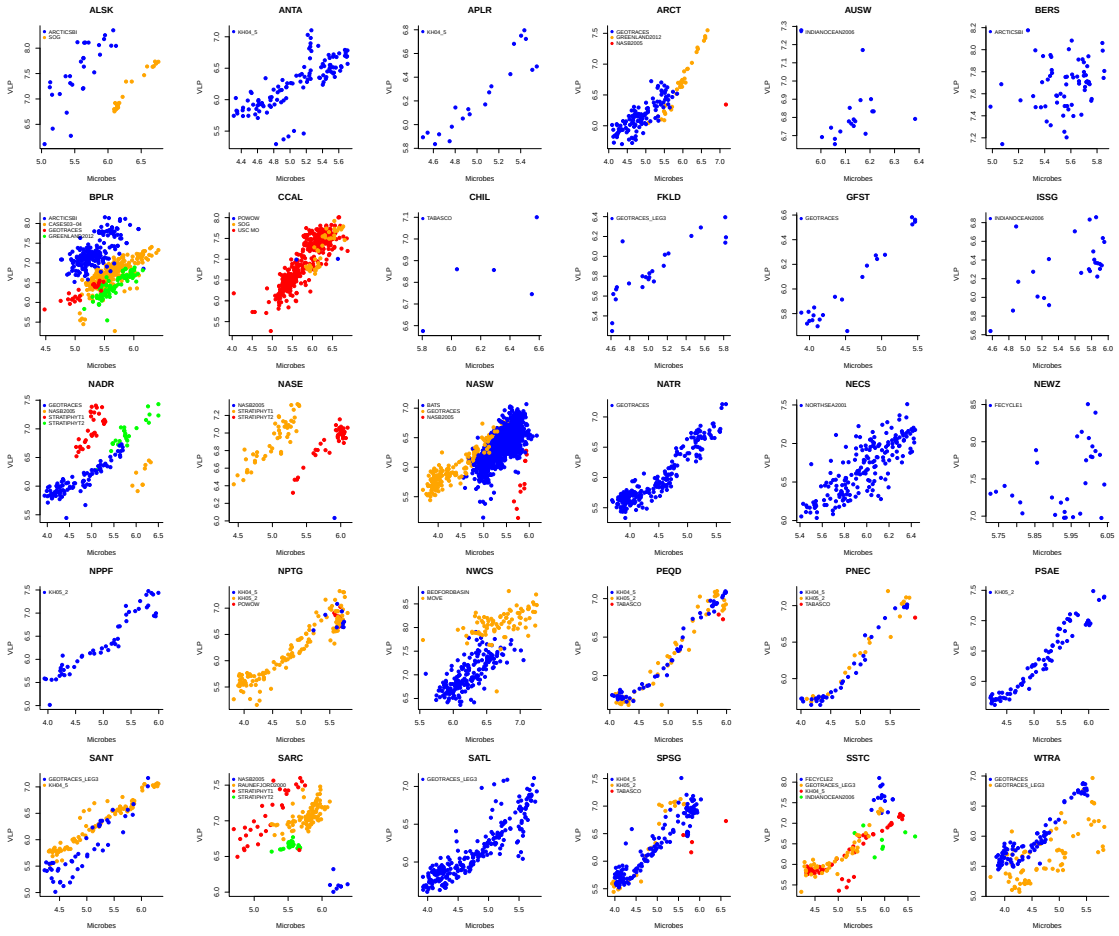


Figure 4.11: **Province data cluster by study** Across the board, Longhurst provinces show study effects virus-microbe space as the data cluster by sampling study.

4.4 Discussion

Viruses are thought the to affect global biogeochemistry by impacting marine microbial growth, productivity, and mortality[96]. Quantifying the relationship between viruses and microbes improves our understanding of this relationship, yet the role environment plays to modulate the interactions between viruses and microbes is unclear. Recent research has shown that there are is a strong effect on viruses by the environment with inroads coming from understanding the relationship between nutrient availability in the ocean and the abundances of viruses and microbes.[98] Here we show that a direct relationship between the measurements of the physical environment and the ratio of viruses to microbes is unclear at global scales by an analysis of environmental covariates observed across 22 studies. Plotting environmental variable values against VMR values from samples provides insight into the complexity of the relationship between environment and VMR. Examining the environmental differences between samples taken one extreme ends of the virus to microbe ratio spectrum suggests three environmental variables (photoactive radiation, temperature, and microbe density) are particularly informative for describing which environmental features may drive differences in VMR ratios. However, a more technical approach was taken via principal component analysis to examine covariation of environmental variables to determine across all virus to micro ratios how environmental features relate to one another and to VMR. Finally by examining the significance of environmental variables when predicting virus density according to study and subsequently mapping the location from which samples were taken to Longhurst provinces to examine samples which are expected to be biochemically similar together, the strength of the environmental signal was assessed at a regional level, ultimately suggesting that variability in VMR may be strongly tied to the studies themselves more so than the environment from which samples were taken. It should be noted that the data does not provide an idea of the viral or microbial composition of samples therefore opening the door to the possibility that high VMR samples may be the result

of measuring one particular viral-host interaction while low VMR samples could be the result of a wholly different virus-host interaction. Likewise, the finding that data clusters by study even within provinces as seen in supplemental figure 4.11 suggests that the opportunity exists to remove the biases in the data coming from processing sample. Some of the bias might be coming from the convenience sampling methods used by collaborators.

4.5 Conclusions

The relationship between VMR and the environment are often limited to descriptions which relate regional physical and nutrient environments as impacting virus and microbe abundances. The analysis here support this finding - the importance of environmental variables is regional at best and does not easily unify to describe global trends in the relationship between environment and VMR. Further, examining the data as cohesive biochemical units by Longhurst provinces did not have the expected effect of controlling for the effect of the environment. In fact virus densities and microbe densities clustered within provinces according to the study through which samples were taken, indicating that the study - be it technician, equipment, or another factor - has a greater impact in determining VMR than the sampling environment, possibly indicating a bias injected in VMR measurements. These findings suggest that an effort to remove systematic bias from the data must be undertaken such that virus and microbe densities can be compared from study to study thus allowing greater precision regarding the effect of the environment on VMR.

4.6 Methods

4.6.1 Data and computing

Data was sampled from 22 studies totaling 5,508 points between 1996 to 2012 and were primarily collected in northern hemisphere during the summer months. The data and R code used in this analysis is archived at [101] as well as being available on the Weitz group

website at through Github at

https://github.com/WeitzGroup/VMR_environmental_covariates.

Analyses were conducted in R version 3.2.3. Packages used for the analyses include RStudio v1.0.143, grid v3.4.0, rmarkdown v1.5, ggplot2 v2.2.1, dplyr v0.5.0, maps v3.1.1[102], mapproj v1.2-4, lme4 v1.1-13, smatr v3.4-3, scales v0.4.1, lmodel2 v1.7-2, knitr v1.16, nlme v3.1-131, reshape2 v1.4.2, lmmfit v1.0, gridExtra v2.2.1, bbmle 1.0.19 [103], and plyr v1.8.4.

The variable PAR_{Depth} describes photoactivethe radiation *PAR* attenuated to the depth at which the sample was taken. *PAR at Depth* was calculated by the quation

$$PAR_{depth} = PAR_{surface} \cdot e^{-.035 \cdot depth} \tag{4.1}$$

where $PAR_{surface}$ is the surface light measured at the sampling location and extinction coefficient is -.035 or roughly a 3.5% reduction in light intensity per meter.

4.6.2 High and low VMR differences and environment

The top 10% and bottom 10% of virus to microbe ratio values were used to create the high VMR and low VMR datasets. Densities for all environmental variables for each of the two data sets were created and overlaid in the same variable space to reveal the different densities observed to high VMR and low VMR samples.

4.6.3 Predicted significance by 23 regression models

23 multivariable regression models were created to predict viral density from the data available in each study, including a dataset with all of the data included. Variables which contained missing data for each study were removed from the model thus some studies had fewer than eight predictive covariates. Possible model covariates include *latitude, longitude, depth, temperature, salinity, chlorophyll-*α*, PAR at Depth*, and *microbe density*.

4.6.4 Principal component analysis

Principal component analysis was conducted on environmental variables normalized by z-transforming each variable such that the analyzed variable had a mean of 0 and a standard deviation of 1.

4.6.5 Mapping Longhurst provinces

The Longhurst Province of each record in the dataset was identified by the Longhurst Province finder script provided by the Chisholm lab's github repository at

https://github.com/thechisholmlab/Longhurst-Province-Finder.

Records which were unable to be fit into a Longhurst province were manually curated such that the nearest Longhurst province was input the record's province.

4.6.6 VMR variability in provinces

The *BBMLE* package was used to fit the province data to the variable-variance and constant variance-models. The special case of the variable-variance model which allows for province-specific parameters was also fit to the data using the BBMLE package.

4.7 Acknowledgments

This work was supported by a grant from the Simons Foundation (SCOPE Award ID 329108 awarded to J.S.W.) and is a contribution of the Simons Collaboration on Ocean Processes and Ecology.

4.8 Supplemental figures

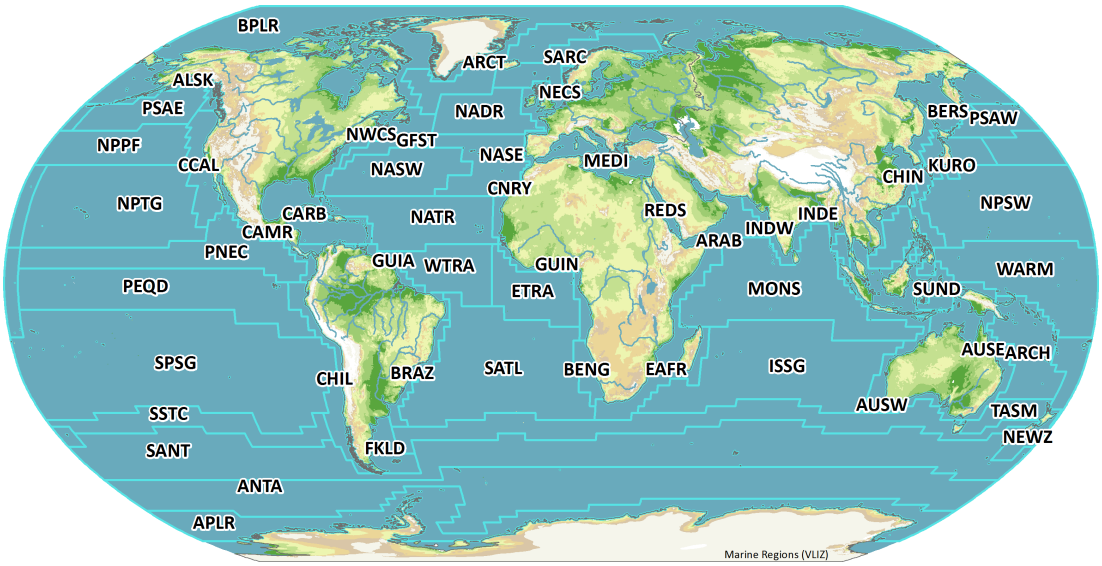


Figure 4.12: **Longhurst provinces delineate ocean areas with similar biochemistry profiles.** Longhurst provinces identify segments of earth's oceans which are biochemically similar. Provinces were sampled unequally in the studies in this analysis. Reprinted from work by Nathalie De Hauwere at Vlaams Instituut voor de Zee (VLIZ).[104]

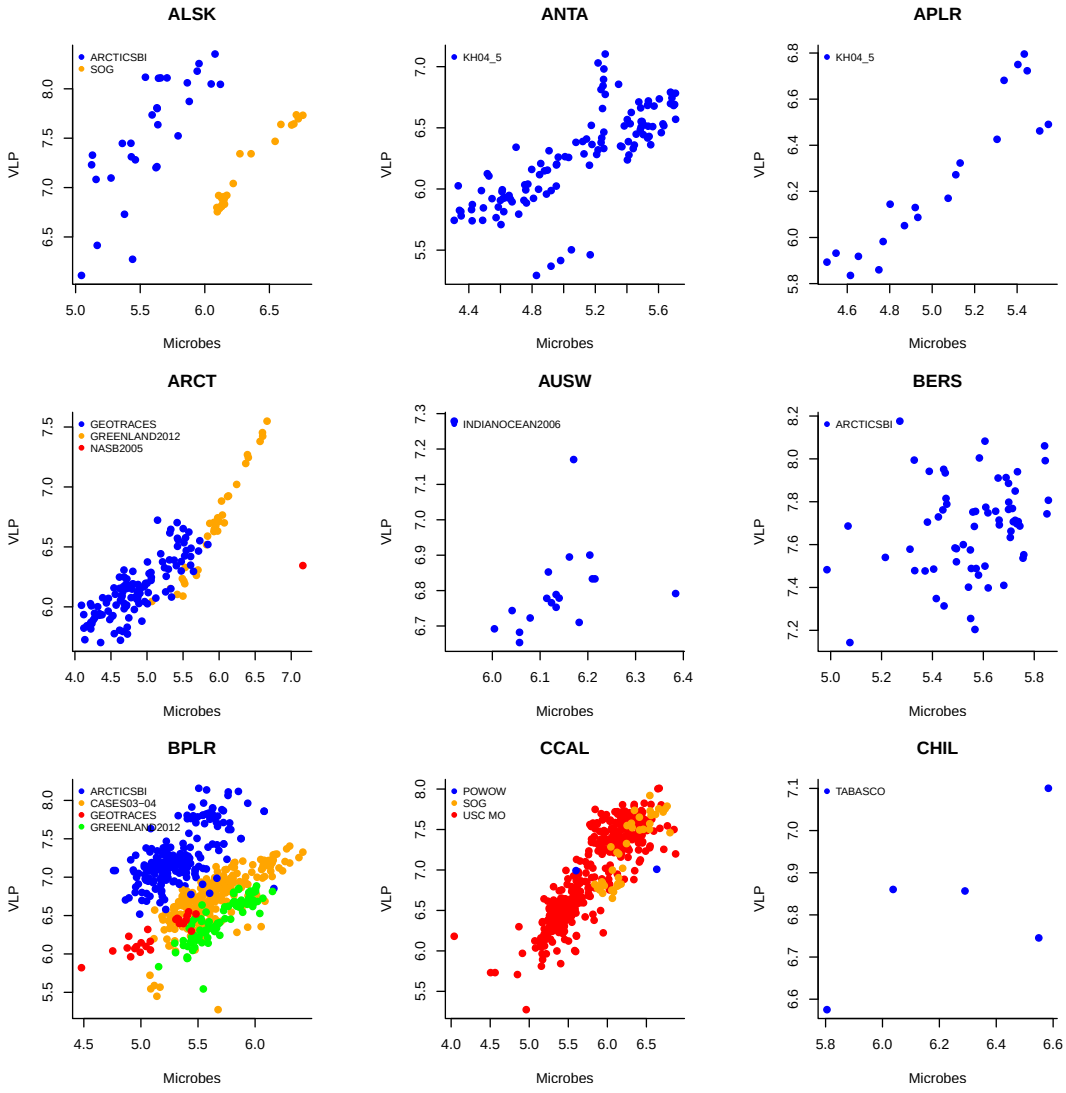


Figure 4.13: **Province data cluster by study (provinces 1-9)** Greater detail shows study effects in Longhurst provinces in virus-microbe space by study.

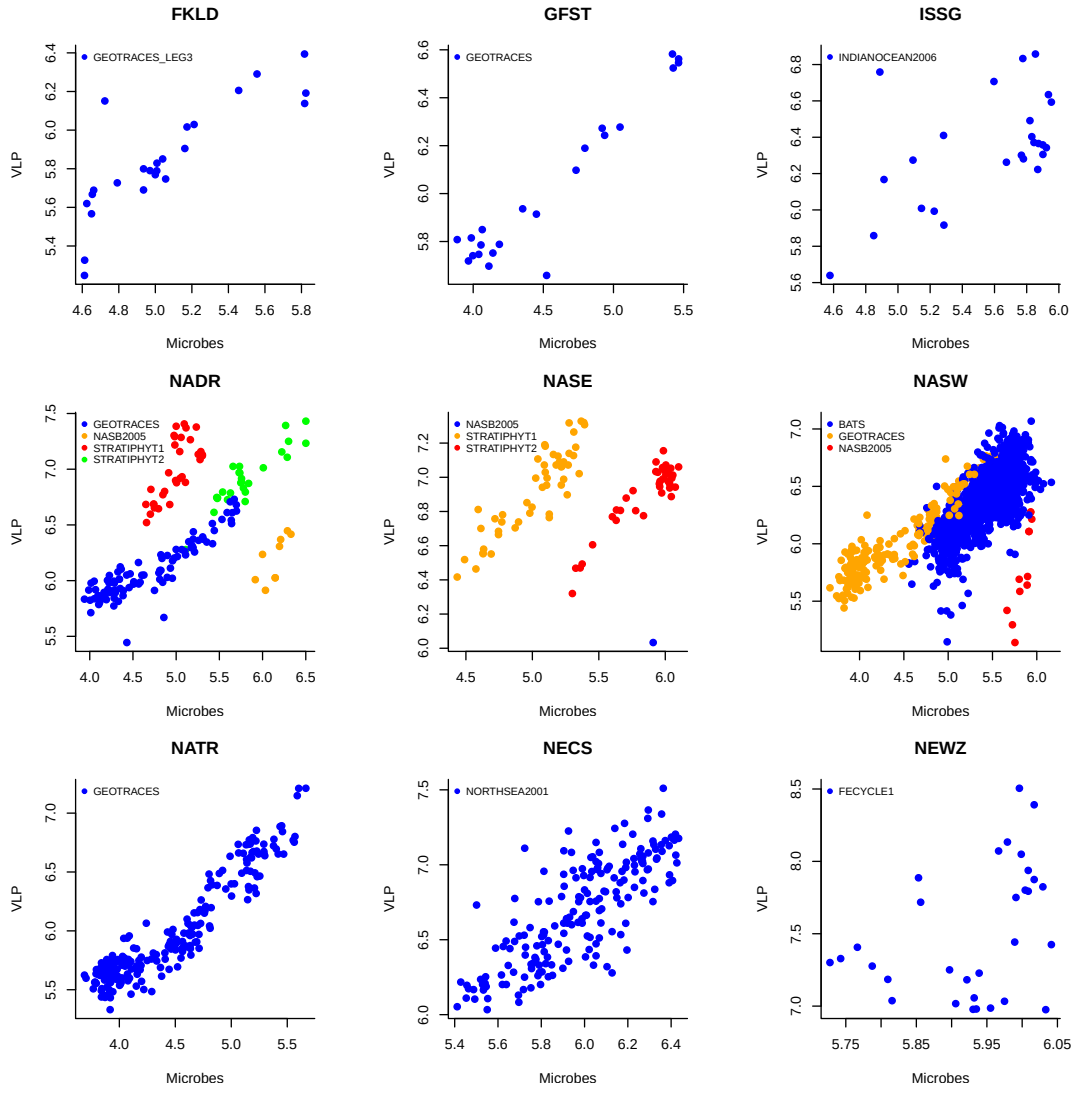


Figure 4.14: **Province data cluster by study (provinces 10-18)** Greater detail shows study effects in Longhurst provinces in virus-microbe space by study.

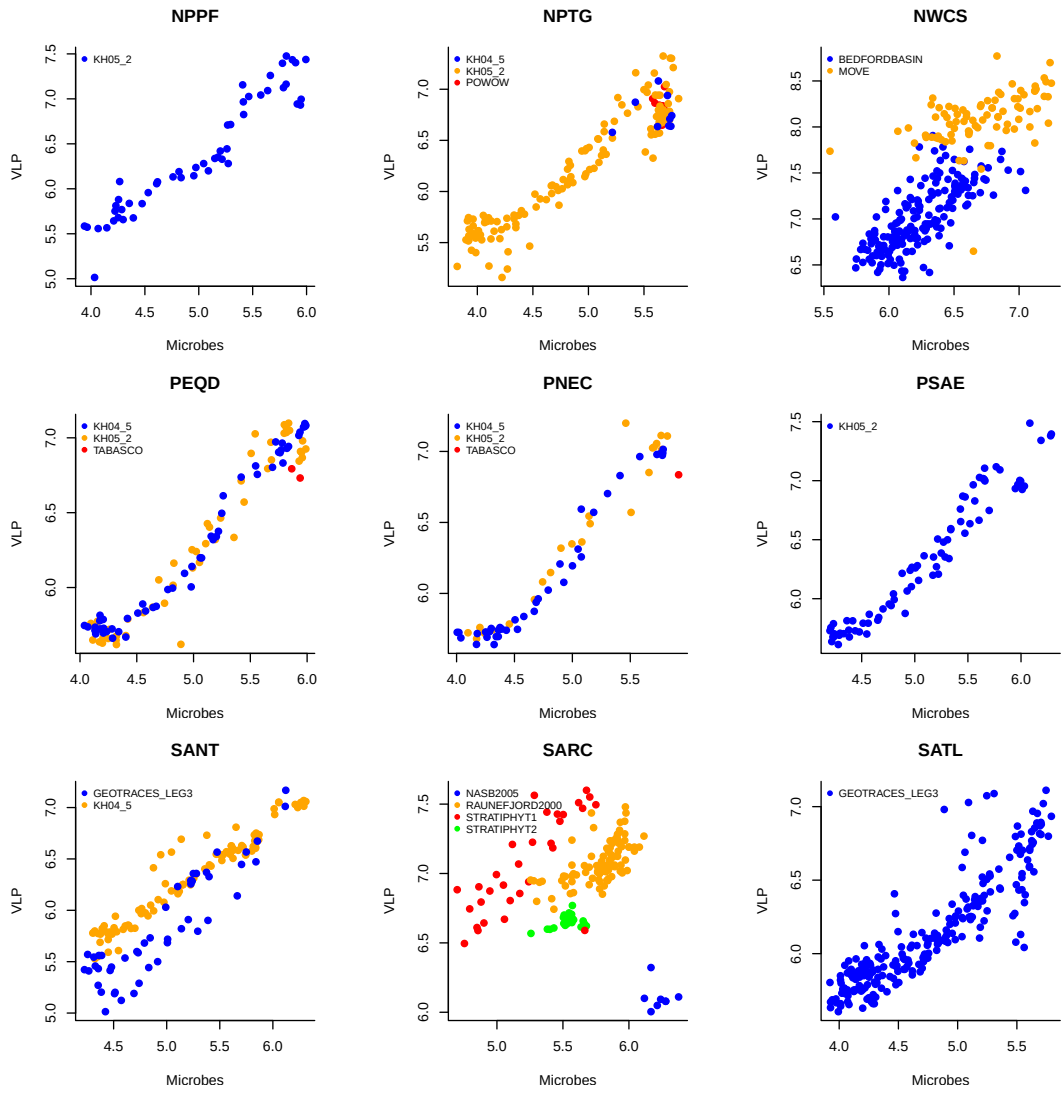


Figure 4.15: **Province data cluster by study (provinces 19-27)** Greater detail shows study effects in Longhurst provinces in virus-microbe space by study.

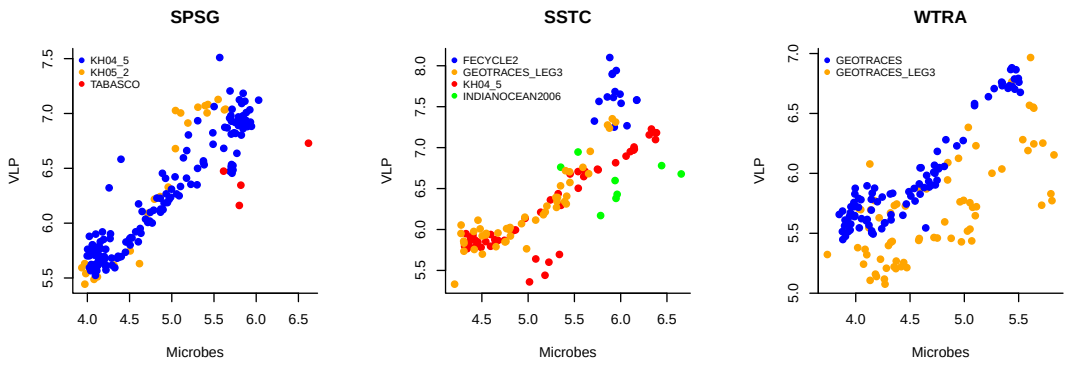


Figure 4.16: **Province data cluster by study (provinces 28, 29, 30)** Greater detail shows study effects in Longhurst provinces in virus-microbe space by study.

CHAPTER 5
A WEB-BASED VISUALIZATION TOOL FOR INVESTIGATING VIRUS-HOST RELATIONSHIPS

5.1 Introduction

Open ocean marine microbe densities are regularly measured between 10^4 and 10^8 viruses ml^{-1} while open ocean marine virus densities are often found between 10^5 and 10^9 microbes ml^{-1}[105]. Understanding the relationship between viruses and hosts is important for gaining insight into drivers of mobile biogeochemistry as viruses are responsible for turning over up to 40% of global ocean organic matter daily[96]. Over time efforts have continued to improve quantifying viruses and microbes densities by imaging. Transmission electron microscopy (TEM) and staining techniques have become the backbone of calculating microbe and virus-like-particle densities. TEM and staining techniques allow for imaging of morphological features and enumerating marine plankton but do little to describe the relationship between viruses and microbes. Static images qualitatively show the relationship between marine hosts and viruses by presence of viruses within microbe cell walls. Likewise, static figures which describe the numerical relationship between host populations and virus populations through time tend to be shown in phase-plane diagrams. Here we provide a visualization toolbox built upon the interactive software Tableau to allow for an interactive experience for users to understand more intuitively the relationship between sampling location, sampling study, and virus and microbe densities. The tableau created here for visualizing the numerical relationship between viruses and microbes shows not only the spatial distribution of sampling sites but a dynamic scatterplot in virus-host space as well as virus-to-microbe ratio distributions for both near-surface and sub-surface samples ($\leq$ 100m and $>$ 100m, respectively). By either selecting the location from which

data was sampled geographically or by selecting specific studies of interest, both the scatter plot as well as the virus-to-microbe ratio histograms are updated to reflect the data selected for visualizing. Such an on-line tool for visualizing the relationship between virus and microbe densities, which can be interrogated through a browser, is not only novel but allows those interested to sidestep the technical requirements typically needed to analyze this kind of data. Therefore this tool should increase collaboration, increase the chances of meaningful insights to come from this data, and aid in hypothesis generation by putting the data at the fingertips of researchers worldwide.

5.2 Results

A tableau was created using all 5,508 records from the Wigington et al 2016 dataset to allow users to interact with the data in real-time. Variables pertinent to the tableau include longitude, latitude, virus density (VLP mL^{-1}), microbe density (microbes mL^{-1}), and the study which collected the sample. This data is available on the left side of a github gh-pages repository which was created to allow for the hosting of the data as well as the tableau. In addition to a public Tableau site which houses all public tableaus, the virus-microbe analysis tableau is found at

http://weitzgroup.github.io/Virus_Microbe_Abundance.

An html page was created to host both downloaded data as well as an iframe which allows the virus-to-microbe ratio tableau to be hosted within the github repository thus allowing for focused directing of visitors to the github page. There are three main figures in the tableau: a map, a scatterplot, and a pair of histograms. A screen shot of the webpage which hosts the tableau and the tableau itself are shown in figure 5.1.

The tableau has built in search tools such as a text entry area which searches the map, zoom (in and out) functions, resetting the zoom of the panel, as well as using the mouse to drag the location of the map where desired. The tableau workbook itself can be downloaded from a link at the bottom right in addition to being able to share the tableau and enter a full-

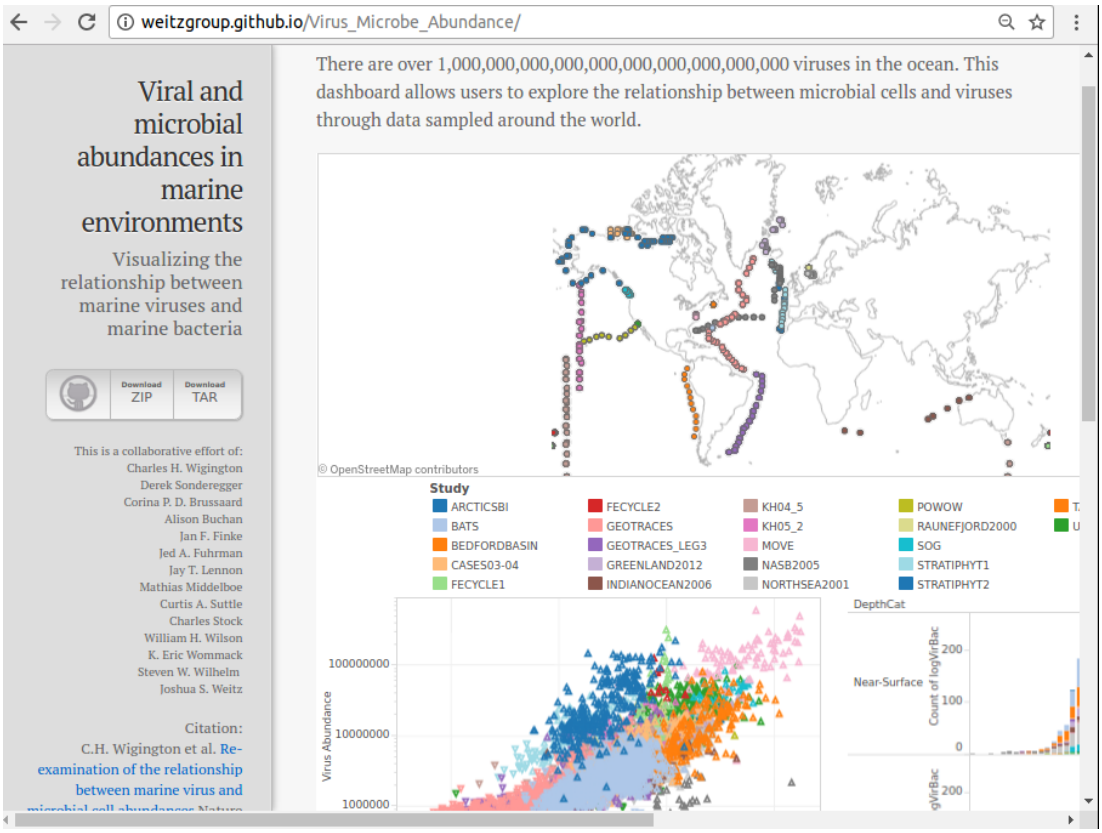


Figure 5.1: **virus-microbe relationship tool hosted on github** The tableau shown here is a three-figure dashboard which displays the data from Wigington et al. 2016 syncronously; filtering the data in any one figure causes the data presented in the other two panels to be filtered.

screen mode. Data selection tools are built into the tableau dashboard to allow for data to be selected using different geometric shapes. The selection of data can be reset, undone, or re-done using the arrows on the bottom left of the tableau iframe. The rectangle, circle, and lasso (a free-form drawing tool to select data), for example, are selected to highlight sampling locations from which observations were collected, this selection will cause a filter to be placed on the data shown in the scatterplot and the histograms such that only the data selected on the map are shown in the scatterplot and histograms. This is shown in figure 5.2. Likewise, selecting observations within the virus and microbe density scatterplot space will update the map and histogram to show where these samples were taken and what their virus to microbe ratios are in the histogram. Finally, selecting the records according to VMR in the histogram populates both the map and the scatterplot with the records which have the selected VMR values.

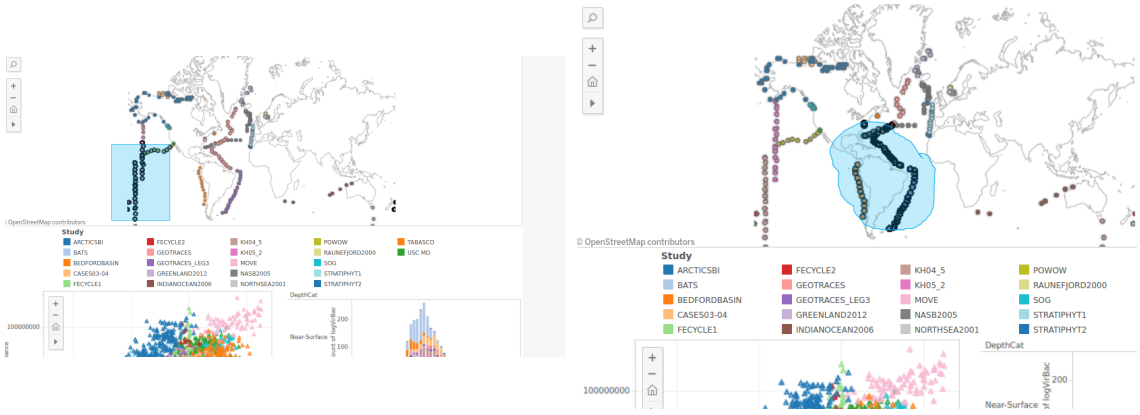


Figure 5.2: **Multiple selection modes are allowed** The tableau displays the data filtered according to selection tools which allow different shapes.

As of Monday, August 14th, there have been 1,686 views.

5.3 Discussion

The availability of an interactive tool which allows marine microbiologist to examine virus-host data interactively reduces the statistical and computing skills needed to analyze the marine virus-host relationship support by data. Here we show it is possible not only to

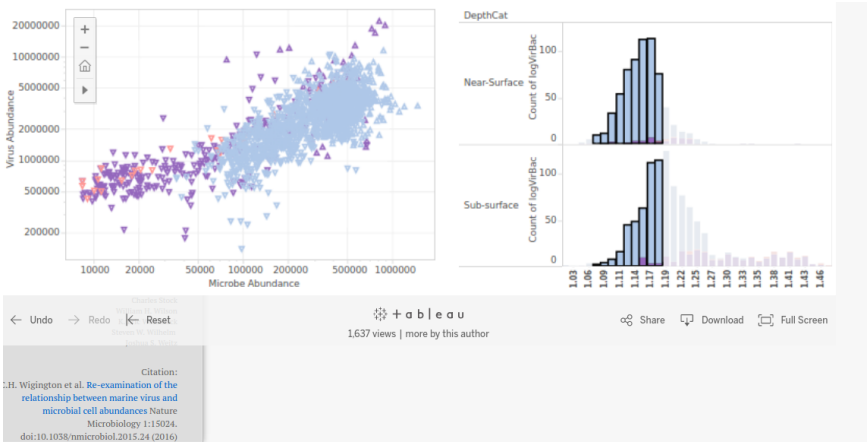


Figure 5.3: **Data can be selected which filters other data spaces** Data can be selected in any of the three panels. Once the data has been selected, a filter is applied to the other two figures in the dashboard such that only the selected data can be presented in the other panels.

identify the location from which samples were taken but to interactively examine the relationship between sampling location and virus to micro ratio. Likewise, the selection of data in virus density and microbe density space which is immediately updated geospatiatially allows for innumerable combinations of analyses to be performed quickly. Further, these combinations of analyses can be created quickly via the data selection tools thus allowing for hypothesis generation and testing almost instantaneously. Finally, such an interactive tool available over the internet allows for hypothesis generation anywhere an internet connection is available.

5.4 Conclusion

The availability of an web-based tool which allows for the interrogation of the relationship between microbes and viruses in marine environments is sure to increase hypothesis generation and data analysis, while reducing the burden of technical computing and statistical needs overhead to generate insights. Further tools such as this one should allow for even greater knowledge of the data as well as even greater knowledge of marine microbiology based on observations of data from the environment. Such quick and intuitive visualiza-

tions of the relationship between marine viruses and their microbial hosts should enable the challenging and testing of long-held beliefs about the relationship between viruses and microbes in marine environments from around the world.

5.5 Methods

5.5.1 Computing environment

The tableau was created through the desktop version of the software Tableau (v9). The workbook for the tableau can be downloaded at

http://weitzgroup.github.io/Virus_Microbe_Abundance.

5.5.2 Tableau Design

The tableau was created in tableau because of the ability to thoroughly integrate crossfiltering across selected data. The crossfiltering ability of tableau allows for a seamless data selection in in the map for example, thus limiting the data shown in the other two visualization spaces. Tableau was selected as the platform for this visualization tool because of its ability to crossfilter relatively large amounts of data - something which could not as easily be accomplished in D3 or javascript.

CHAPTER 6

CONCLUSIONS

Marine microbes are present in unfathomable numbers in Earth's ocean, estimated to be $\sim 10^{30}$ in number while viruses are even more abundant or $\sim 10^{31}$. This thesis examined the 10 to 1 ratio through the use of data collected from around the world to quantify this model as it describes the relationship between marine viruses and their microbial hosts. This 10:1 was shown to be insufficient for describing viral density relative to microbial density, having negative R^2 values for both the near surface and sub-surface data. Furthermore, the width of the distribution of virus to microbe ratios in both the near-surface and subsurface samples showed that the 10:1 is a poor descriptor and described the less than 5% of the data collected. The power-law model proposed is a marked improvement to the 10:1 ratio, leading to the conclusion that the relationship between viruses and microbes in marine environments is non only non-linear, but sub-linear meaning that with increasing microbial density, virus density increases however the ratio of viruses to microbes decreases with increasing microbe density.

The variability in the ratio of viruses to microbes across microbe densities indicates that samples taken from the sub-surface have lower variability than the near-surface and that this increase in variability with microbe density is real and is a result of differences VMR offset by study, not because of particularly wide bands of variability observed at high microbial densities. This observed effect was examined with respect to differences in the environment which did not show any particular environmental variable as the driver of VMR values but instead is the result of combinations of enrvionmental variables. This unclear description of VMR according to environment is highlighted by the changing importance of environemntal variables across studies.

Even though Longhurst provinces were invoked to cluster samples according to similar

environments, study effects were nevertheless observed across provinces which were sampled by multiple studies. This difference in VMR offset according to study, where records within the same province should be influenced identically by the environment, indicate that differences in studies in terms of methods, equipment, or even technician are impacting VMR ratios and thus muddying our understanding of what environmental factors are driving VMR.

To shortcut the overhead required to analyze this data set, an online tableau was created to for easily examining the data interatively. The data is linked across three visualizations, a global sampling location map, a scatterplot relating microbial density and viral density, and two histograms which describe the densities VMR values analyzed. The ability to generate hypothesis around observations from this tableau allow for greater visibility of the data as well as interrogation of the data by new and experienced marine microbiologists alike.

REFERENCES

[1] T. Twort, "An investigation on the nature of ultra-microscopic viruses," *Lancet*, vol. 2, pp. 1241–1243, 1915.

[2] F. d'Hérelle, "Sur un microbe invisible antagoniste des bacilles dysentèriques," *Cr. R. Acad. Sci. Paris*, vol. 165, 1917.

[3] A. Kriss and E. Rukina, *Bacteriophage in the sea. (In Russian)*. Dok Akad Nauk SSSR, 1947, vol. 57.

[4] F. Torrella and R. Morita, "Evidence by electron-micrographs for a high-incidence of bacteriophage particles in the waters of Yaquina Bay, Oregon - ecological and taxonomical implications," *Applied and Environmental Microbiology*, vol. 37, no. 4, pp. 774–778, 1979.

[5] O. Bergh, K. Y. Borsheim, G. Bratbak, and M. Heldal, "High abundance of viruses found in aquatic environments," *Nature*, vol. 340, no. 6233, pp. 467–468, 1989.

[6] V. Racaniello, *Http://www.virology.ws/2009/07/06/detecting-viruses-the-plaque-assay/*.

[7] J. A. Fuhrman, "Marine viruses and their biogeochemical and ecological effects.," *Nature*, vol. 399, pp. 541–8, Jun. 1999.

[8] Ø. Bergh, K. Y. BØrsheim, G. Bratbak, and M. Heldal, "High abundance of viruses found in aquatic environments," *Nature*, vol. 340, no. 6233, pp. 467–468, 1989.

[9] J. S. Weitz, *Quantitative viral ecology: dynamics of viruses and their microbial hosts*. Princeton University Press, 2016.

[10] T. Thingstad, "Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems," *Limnology and Oceanography*, vol. 45, pp. 1320–1328, 2000.

[11] B Knowles, C. Silveira, B. Bailey, K Barott, V. Cantu, A. Cobián-Güemes, F. Coutinho, E. Dinsdale, B Felts, K. Furby, *et al.*, "Lytic to temperate switching of viral communities," *Nature*, vol. 531, no. 7595, pp. 466–470, 2016.

[12] K. J. Parikka, M. Le Romancer, N. Wauters, and S. Jacquet, "Deciphering the virus-to-prokaryote ratio (vpr): Insights into virus–host relationships in a variety of ecosystems," *Biological Reviews*, vol. 92, no. 2, pp. 1081–1100, 2017.

[13] J. Weitz, S. J. Beckett, J. R. Brum, B. Cael, and J. Dushoff, "Lysis, lysogeny, and virus-microbe ratios," *Biorxiv*, p. 051 805, 2016.

[14] C. Suttle, "Viruses in the sea," *Nature*, vol. 437, pp. 356–361, 2005.

[15] C. P. D. Brussaard, S. W. Wilhelm, F. Thingstad, M. G. Weinbauer, G. Bratbak, M. Heldal, S. A. Kimmance, M. Middelboe, K. Nagasaki, J. H. Paul, D. C. Schroeder, C. A. Suttle, D. Vaque, and K. E. Wommack, "Global-scale processes with a nanoscale drive: the role of marine viruses," *ISME Journal*, vol. 2, 575–578, 2008.

[16] F. Rohwer and R. Thurber, "Viruses manipulate the marine environment," *Nature*, vol. 459, pp. 207–212, 2009.

[17] J. Weitz and S. Wilhelm, "Ocean viruses and their effects on microbial communities and biogeochemical cycles," *F1000 biology reports*, vol. 4, p. 17, 2012.

[18] L. Jover, T. Effler, A. Buchan, S. Wilhelm, and J. Weitz, "The elemental composition of virus particles: Implications for marine biogeochemical cycles," *Nature Reviews Microbiology*, vol. 12, pp. 519–528, 2014.

[19] J. Weitz, H. Hartman, and S. Levin, "Coevolutionary arms races between bacteria and bacteriophage," *Proceedings of the National Academy of Sciences, USA*, vol. 102, pp. 9535–9540, 2005.

[20] S. Avrani, D. Schwartz, and D. Lindell, "Virus-host swinging party in the oceans: Incorporating biological complexity into paradigms of antagonistic coexistence.," *Mobile Genetic Elements*, vol. 2, pp. 88–95, 2012.

[21] J. Payet and C. Suttle, "To kill or not to kill : The balance between lytic and lysogenic viral infection is driven by trophic status," *Limnology and Oceanography*, vol. 58, pp. 465–474, 2013.

[22] A. Murray and G. Jackson, "Viral dynamics: A model of the effects of size, shape, motion and abundance of single-celled planktonic organisms and other particles," *Marine Ecology Progress Series*, vol. 89, pp. 103–116, 1992.

[23] R. Maranger and D. Bird, "Viral abundance in aquatic systems - a comparison between marine and fresh-waters," *Marine Ecology Progress Series*, vol. 121, no. 1-3, pp. 217–226, 1995.

[24] K. E. Wommack and R. R. Colwell, "Virioplankton: Viruses in aquatic ecosystems," *Microbiology and Molecular Biology Reviews*, vol. 64, no. 1, pp. 69–114, 2000.

[25] M. Weinbauer, "Ecology of prokaryotic viruses," *FEMS Microbiology Reviews*, vol. 28, no. 2, pp. 127–181, 2004.

[26] K. Mojica, W. van de Poll, M. Kehoe, J. Huisman, K. Timmermans, A. Buma, H. van der Woerd, L. Hahn-Woernle, H. Dijkstra, and C. Brussaard, "Phytoplankton community structure in relation to vertical stratification along a north-south gradient in the Northeast Atlantic Ocean," *Limnology and Oceanography*, 2015.

[27] R. Edwards and F. Rohwer, "Viral metagenomics," *Nat. Rev. Microb.*, vol. 3, pp. 504–510, 2005.

[28] J. Brum, B. Hurwitz, O. Schofield, H. Ducklow, and M. Sullivan, "Seasonal time bombs: Dominant temperate viruses affect southern ocean microbial dynamics," *ISME J*, Aug. 2015.

[29] J. Weitz, T. Poisot, J. Meyer, C. Flores, M. Valverde S.and Sullivan, and M. Hochberg, "Phage-bacteria infection networks," *Trends in Microbiology*, vol. 21, pp. 82–91, 2013.

[30] R. Danovaro, C. Corinaldesi, A. Dell'Anno, J. Fuhrman, J. Middelburg, R. Noble, and C. Suttle, "Marine viruses and global climate change," *FEMS Microbiology Reviews*, vol. 35, no. 6, pp. 993–1034, 2011.

[31] A. Morel and J. Berthon, "Surface pigments, algal biomass profiles, and potential production of the euphotic layer: Relationships reinvestigated in view of remote-sensing applications," *Limnology and Oceanography*, vol. 34, no. 8, pp. 1545–1562, 1989.

[32] R. J. Parsons, M. Breitbart, M. W. Lomas, and C. A. Carlson, "Ocean time-series reveals recurring seasonal patterns of virioplankton dynamics in the northwestern Sargasso Sea," *The ISME Journal*, vol. 6, pp. 273–284, 2011.

[33] J. A. Fuhrman, I. Hewson, M. S. Schwalbach, J. A. Steele, M. V. Brown, and S. Naeem, "Annually reoccurring bacterial communities are predictable from ocean conditions," *Proceedings of the National Academy of Sciences USA*, vol. 103, pp. 13 104–13 109, 2006.

[34] D. De Corte, E. Sintes, T. Yokokawa, T. Reinthaler, and G. J. Herndl, "Links between viruses and prokaryotes throughout the water column along a North Atlantic latitudinal transect," *The ISME Journal*, vol. 6, no. 8, pp. 1566–1577, 2012.

[35] W. K. W. Li and P. M. Dickie, "Monitoring phytoplankton, bacterioplankton, and virioplankton in a coastal inlet (bedford basin) by flow cytometry," *Cytometry*, vol. 44, no. 3, pp. 236–246, 2001.

[36] Y. Yang, T. Yokokawa, C. Motegi, and T. Nagata, "Large-scale distribution of viruses in deep waters of the Pacific and Southern oceans," *Aquatic Microbial Ecology*, vol. 71, no. 3, pp. 193–202, 2013.

[37] J. L. Clasen, S. M. Brigden, J. P. Payet, and C. A. Suttle, "Evidence that viral abundance across oceans and lakes is driven by different biological factors," *Freshwater Biology*, vol. 53, no. 6, pp. 1090–1100, 2008.

[38] A. Balsom, *Macroinfaunal community composition and biomass, and bacterial and viral abundances from the gulf of alaska to the canadian archipelago: A biodiversity study*, Electronic Book, 2003.

[39] R. F. Strzepek, M. T. Maldonado, J. L. Higgins, J. Hall, K. Safi, S. W. Wilhelm, and P. W. Boyd, "Spinning the "ferrous wheel": The importance of the microbial community in an iron budget during the fecycle experiment," *Global Biogeochemical Cycles*, vol. 19, no. 4, 2005.

[40] A. R. Matteson, S. N. Loar, S. Pickmere, J. M. DeBruyn, M. J. Ellwood, P. W. Boyd, D. A. Hutchins, and S. W. Wilhelm, "Production of viruses during a spring phytoplankton bloom in the south pacific ocean near of new zealand," *FEMS Microbiology Ecology*, vol. 79, no. 3, pp. 709–719, 2012.

[41] J. M. Rowe, M. A. Saxton, M. T. Cottrell, J. M. DeBruyn, G Mine Berg, D. L. Kirchman, D. A. Hutchins, and S. W. Wilhelm, "Constraints on viral production in the Sargasso Sea and North Atlantic," *Aquatic Microbial Ecology*, vol. 52, no. 3, pp. 233–244, 2008.

[42] S. W. Wilhelm, W. H. Jeffrey, A. L. Dean, J. Meador, J. D. Pakulski, and D. L. Mitchell, "Uv radiation induced dna damage in marine viruses along a latitudinal gradient in the southeastern pacific ocean," *Aquatic Microbial Ecology*, vol. 31, no. 1, pp. 1–8, 2003, 652AR Times Cited:39 Cited References Count:33.

[43] K. Wang, K. E. Wommack, and F. Chen, "Abundance and distribution of *synechococcus* spp. and cyanophages in the Chesapeake Bay," *Applied and Environmental Microbiology*, vol. 77, no. 21, pp. 7459–7468, 2011.

[44] C. Suttle, "Marine viruses - major players in the global ecosystem," *Nature Reviews Microbiology*, vol. 5, 801–812, 2007.

[45] R. Danovaro, A. Dell'Anno, C. Corinaldesi, M. Magagnini, R. Noble, C. Tamburini, and M. Weinbauer, "Major viral impact on the functioning of benthic deep-sea ecosystems," *Nature*, vol. 454, 1084–U27, 2008.

[46] J. Brum and M. Sullivan, "Rising to the challenge: Accelerated pace of discovery transforms marine virology," *Nat Rev Micro*, vol. 13, no. 3, pp. 147–159, Mar. 2015.

[47] J. Weitz, C. Stock, S. Wilhelm, L. Bourouiba, M. Coleman, A. Buchan, M. Follows, J. Fuhrman, L. Jover, J. Lennon, M. Middelboe, D. Sonderegger, C. Suttle, B. Taylor, T. Thingstad, W. Wilson, and K. Wommack, "A multitrophic model to quantify the effects of marine viruses on microbial food webs and ecosystem processes," *ISME J*, vol. 9, no. 6, pp. 1352–1364, Jun. 2015.

[48] C. Suttle and A. Chan, "Marine cyanophages infecting oceanic and coastal strains of *Synechococcus*: abundance, morphology, cross-infectivity and growth characteristics," *Marine Ecology Progress Series*, vol. 92, pp. 99–109, 1993.

[49] M. De Paepe and F. Taddei, "Viruses' life history: Towards a mechanistic basis of a trade-off between survival and reproduction among phages," *PLoS Biology*, vol. 4, e193, 2006.

[50] T. Thingstad and R. Lignell, "Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand," *Aquatic Microbial Ecology*, vol. 13, pp. 19–27, 1997.

[51] S. Giovannoni, B. Tempterton, and Y. Zhao, "Giovannoni et al. reply, "SAR11 viruses and defensive host strains"," *Nature*, vol. 499, E4–E5, 2013.

[52] C. Carreira, M. Larsen, R. Glud, C. Brussaard, and M. Middelboe, "Heterogeneous distribution of prokaryotes and viruses at the microscale in a tidal sediment," *Aquatoc Microbial Ecololgy*, vol. 69, pp. 183–192, 2013.

[53] G. Bratbak and M. Heldal, "Viruses-the new players in the game: Their ecological role and could they mediate genetic exchange by transduction?" In *Molecular Ecology of Aquatic Microbes*, I. Joint, Ed., vol. 238, Berlin, Germany: Springer-Verlag KG, 1995, pp. 249–264.

[54] S. Williamson, L. Houchin, L. McDaniel, and J. Paul, "Seasonal variation in lysogeny as depicted by prophage induction in Tampa Bay, Florida," *Appl Env Micro*, vol. 68, pp. 4307–4314, 2002.

[55] I. Hatton, K. McCann, J. Fryxell, T. Davies, M. Smerlak, A. Sinclair, and M. Loreau, "The predator-prey power law: Biomass scaling across terrestrial and aquatic biomes," *Science*, vol. 349, no. 6252, 2015.

[56] C. Suttle and J. Fuhrman, "Enumeration of virus particles in aquatic or sediment samples by epifluorescence microscopy," in *Manual of Aquatic Viral Ecology*, S. W.

Wilhelm, M. G. Weinbauer, and C. Suttle, Eds., Waco, TX: American Society of Limnology and Oceanography, 2010, pp. 145–153.

[57] C. Brussaard, J. Payet, C. Winter, and M. Weinbauer, "Quantification of aquatic viruses by flow cytometry," in *Manual of Aquatic Viral Ecology*, S. Wilhelm and M. Weinbauer, Eds., 2010, pp. 102–109.

[58] Y. Tomaru and K. Nagasaki, "Flow cytometric detection and enumeration of dna and rna viruses infecting marine eukaryotic microalgae," *Journal of Oceanography*, vol. 63, no. 2, pp. 215–221, 2007.

[59] J. Labonté and C. Suttle, "Previously unknown and highly divergent ssdna viruses populate the oceans," *The ISME journal*, vol. 7, no. 11, pp. 2169–2177, 2013.

[60] C. Brussaard, D. Marie, and G. Bratbak, "Flow cytometric detection of viruses," *Journal of Virological Methods*, vol. 85, pp. 175 –182, 2000.

[61] G. Steward, A. Culley, J. Mueller, E. Wood-charlson, M. Belcaid, and G. Poisson, "Are we missing half of the viruses in the ocean ?" *ISME J.*, vol. 7, pp. 672–679, 2012.

[62] D. Raoult, S. Audic, C. Robert, C. Abergel, P. Renesto, H. Ogata, B. L. Scola, M. Suzan, and J. Claverie, "The 1.2-megabase genome sequence of mimivirus," *Science*, vol. 306, pp. 1344–1350, 2004.

[63] A. Certes, "Sur la culture, à labri des germes atmosphériques, des eaux et des sédiments rapportés par les expeditions du travailleur et du talisman 1882 - 1883," *Comptes Rendus Acad Sci*, no. 98, p. 4, 1884.

[64] B. Fischer, "Bakteriologische untersuchungen auf einer reise nach west- indien," *Zeitschrift für Hygiene und Infektionskrankheiten*, no. 1, p. 44, 1886.

[65] H. Russell, "Untersuchungen ber im golf von neapel lebende bakterien," *Zeitschrift für Hygiene und Infektionskrankheiten*, vol. 11, p. 42, 1891.

[66] B. Fischer, *Die Bakterien des Meeres nach den Untersuchungen der Plankton-Expedition unter gleichzeitiger Bercksichtigung einiger älterer und neuerer Untersuchungen*. Keil: Lipsius & Tischer, 1894.

[67] F. Arloing and Chavanne, "On the influence of the environment on the bacteriophage, electrolytes and the concentration of H ions," *Comptes Rendus Des Séances De La Société De Biologie Et De Ses Filiales*, vol. 93, pp. 531–532, 1925.

[68] C. ZoBell, *Marine microbiology, a monograph on hydrobacteriology*, ser. Marine microbiology. A monograph on hydrobacteriology. 1946, xv, 1 , 240 p.

[69] A. Carlucci and D. Pramer, "An evaluation of factors affecting the survival of Escherichia-coli in sea water," *Applied Microbiology*, vol. 8, no. 4, pp. 251–254, 1960.

[70] L. Proctor and J. A. Fuhrman, "Viral mortality of marine-bacteria and cyanobacteria," *Nature*, vol. 343, no. 6253, pp. 60–62, 1990.

[71] J. Fuhrman and C. Suttle, "Viruses in marine planktonic systems," *Oceanography*, vol. 6, no. 2, pp. 51–63, 1993.

[72] S. Hara, K. Terauchi, and I. Koike, "Abundance of viruses in marine waters - assessment by epifluorescence and transmission electron-microscopy," *Applied and Environmental Microbiology*, vol. 57, no. 9, pp. 2731–2734, 1991.

[73] J. Paul, S. C. Jiang, and J. B. Rose, "Concentration of viruses and dissolved DNA from aquatic environments by vortex flow filtration," *Applied and Environmental Microbiology*, vol. 57, no. 8, pp. 2197–2204, 1991.

[74] K. Wommack, R. Hill, M. Kessel, E. Russekcohen, and R. Colwell, "Distribution of viruses in the Chesapeake Bay," *Applied and Environmental Microbiology*, vol. 58, no. 9, pp. 2965–2970, 1992.

[75] W. Cochlan, J. Wikner, G. Steward, D. Smith, and F. Azam, "Spatial-distribution of viruses, bacteria and chlorophyll-a in neritic, oceanic and estuarine environments," *Marine Ecology Progress Series*, vol. 92, no. 1-2, pp. 77–87, 1993.

[76] J. Paul, J. B. Rose, S. C. Jiang, C. A. Kellogg, and L. Dickson, "Distribution of viral abundance in the reef environment of Key Largo, Florida," *Applied and Environmental Microbiology*, vol. 59, no. 3, pp. 718–724, 1993.

[77] G. Bratbak, M. Heldal, S. Norland, and T. Thingstad, "Viruses as partners in spring bloom microbial trophodynamics," *Applied and Environmental Microbiology*, vol. 56, no. 5, pp. 1400–1405, 1990.

[78] M. Heldal and G. Bratbak, "Production and decay of viruses in aquatic environments," *Marine Ecology Progress Series*, vol. 72, pp. 205–212, 1991.

[79] D. Smith, G. Steward, F. Azam, and J. Hollibaugh, "Virus and bacteria abundance in the Drake Passage during January and August 1991," *Antarctic Journal of the United States*, vol. 27, pp. 125–127. 1992.

[80] J. Boehme, M. Frischer, S. Jiang, C. Kellogg, S. Pichard, J. Rose, C. Steinway, and J. Paul, "Viruses, bacterioplankton, and phytoplankton in the southeastern Gulf-of-Mexico - distribution and contribution to oceanic DNA pools," *Marine Ecology Progress Series*, vol. 97, no. 1, pp. 1–10, 1993.

[81] M. Weinbauer, D. Fuks, and P. Peduzzi, "Distribution of viruses and dissolved DNA along a coastal trophic gradient in the Northern Adriatic Sea," *Applied and Environmental Microbiology*, vol. 59, no. 12, pp. 4074–4082, 1993.

[82] S. Jiang and J. Paul, "Seasonal and diel abundance of viruses and occurrence of lysogeny/bacteriocinogeny in the marine-environment," *Marine Ecology Progress Series*, vol. 104, no. 1-2, pp. 163–172, 1994.

[83] S. Chibani-Chennoufi, A. Bruttin, M. Dillmann, and H. Brussow, "Phage-host interaction: An ecological perspective," *Journal of Bacteriology*, vol. 186, no. 12, pp. 3677–3686, 2004.

[84] S. W. Wilhelm and A. R. Matteson, "Freshwater and marine virioplankton: A brief overview of commonalities and differences," *Freshwater Biology*, vol. 53, no. 6, pp. 1076–1089, 2008.

[85] R. Milo, P. Jorgensen, U. Moran, G. Weber, and M. Springer, "BioNumbers-the database of key numbers in molecular and cell biology," *Nucleic Acids Research*, vol. 38, pp. D750–D753, 2010.

[86] S. W. Wilhelm and C. A. Suttle, "Viruses and nutrient cycles in the sea," *BioScience*, vol. 49, pp. 781–788, 1999.

[87] E. Buitenhuis, M Vogt, R Moriarty, N Bednarsek, S. Doney, K. Leblanc, C Le Quéré, Y.-W. Luo, C O'Brien, T O'Brien, *et al.*, "Maredat: Towards a world atlas of marine ecosystem data," *Earth System Science Data*, vol. 5, no. 2, p. 227, 2013.

[88] C. Suttle, "Viruses in the sea," *Nature*, vol. 437, pp. 356–361, 2005.

[89] P Landschützer, N Gruber, D. Bakker, and U Schuster, "Recent variability of the global ocean carbon sink," *Global Biogeochemical Cycles*, vol. 28, no. 9, pp. 927–949, 2014.

[90] S. W. Wilhelm and C. A. Suttle, "Viruses and nutrient cycles in the sea viruses play critical roles in the structure and function of aquatic food webs," *Bioscience*, vol. 49, no. 10, pp. 781–788, 1999.

[91] K. E. Wommack and R. R. Colwell, "Virioplankton: Viruses in aquatic ecosystems," *Microbiology and molecular biology reviews*, vol. 64, no. 1, pp. 69–114, 2000.

[92] S. C. Jiang and J. H. Paul, "Seasonal and diel abundance of viruses and occurrence of lysogeny/bacteriocinogeny in the marine environment," *Marine Ecology-Progress Series*, vol. 104, p. 163, 1994.

[93] L. A. Drake, K.-H. Choi, A. E. Haskell, and F. C. Dobbs, "Vertical profiles of virus-like particles and bacteria in the water column and sediments of chesapeake bay, usa," *Aquatic Microbial Ecology*, vol. 16, no. 1, pp. 17–25, 1998.

[94] C. H. Wigington, D. Sonderegger, C. P. Brussaard, A. Buchan, J. F. Finke, J. A. Fuhrman, J. T. Lennon, M. Middelboe, C. A. Suttle, C. Stock, *et al.*, "Re-examination of the relationship between marine virus and microbial cell abundances," *Nature microbiology*, vol. 1, p. 15 024, 2016.

[95] C. H. Wigington, "VMR_variability," 2017.

[96] C. A. Suttle, "Marine virusesmajor players in the global ecosystem," *Nature Reviews Microbiology*, vol. 5, no. 10, pp. 801–812, 2007.

[97] S. C. Doney, M. Ruckelshaus, J. E. Duffy, J. P. Barry, F. Chan, C. A. English, H. M. Galindo, J. M. Grebmeier, A. B. Hollowed, N. Knowlton, *et al.*, "Climate change impacts on marine ecosystems," 2011.

[98] J. F. Finke, B. P. Hunt, C. Winter, E. C. Carmack, and C. A. Suttle, "Nutrients and other environmental factors influence virus abundances across oxic and hypoxic marine environments," *Viruses*, vol. 9, no. 6, p. 152, 2017.

[99] A. S. Cabral, M. M. Lessa, P. C. Junger, F. L. Thompson, and R. Paranhos, "Virioplankton dynamics are related to eutrophication levels in a tropical urbanized bay," *PloS one*, vol. 12, no. 3, e0174653, 2017.

[100] A. Longhurst, S. Sathyendranath, T. Platt, and C. Caverhill, "An estimate of global primary production in the ocean from satellite radiometer data," *Journal of plankton Research*, vol. 17, no. 6, pp. 1245–1271, 1995.

[101] C. H. Wigington, "VMR_environmental_covariates," 2017.

[102] O. S. code by Richard A. Becker, A. R.W. R. version by Ray Brownrigg. Enhancements by Thomas P Minka, and A. Deckmyn., *Maps: Draw geographical maps*, R package version 3.1.1, 2016.

[103] B. Bolker and R. D. C. Team, *Bbmle: Tools for general maximum likelihood estimation*, R package version 1.0.19, 2017.

[104] S. Claus, P. Deckers, B. Vanhoorne, F. Hernandez, and E. V. Berghe, "Developments and geographic interface of the vlimar gazetteer," 2006.

[105] J. A. Fuhrman, "Marine viruses and their biogeochemical and ecological effects," *Nature*, vol. 399, no. 6736, p. 541, 1999.